CHAPTER

4

Hierarchical *k*-Means: A Hybrid Clustering Algorithm and Its Application to Study Gene Expression in Lung Adenocarcinoma

Mohammad Shabbir Hasan and Zhong-Hui Duan

Department of Computer Science, College of Arts and Sciences, University of Akron, Akron, USA

1 INTRODUCTION

Gene products such as proteins or RNA are created from the inheritable information contained in a gene (Hunter and Holm, 1992). Traditional molecular biology focuses on studying individual genes in isolation for determining gene functions. However, it is not suitable for determining complex gene interactions or for explaining the nature of complex biological processes due to the large number of genes. For this purpose, examining the expression pattern of a large number of genes in parallel is required (Michaels *et al*., 1998). With the advancement of large-scale transcription profiling technology, DNA microarrays have become a useful tool that allows the analysis of the gene expression pattern at the genome level (Gresham *et al*., 2008). In genetic-mapping studies, DNA microarrays have been widely used on polymorphisms between parental genotypes and have facilitated the discovery of gene expression markers (Gresham *et al*., 2008; Wang *et al*., 2009). Due to its importance, efficient algorithms are necessary to analyze the DNA microarray data set accurately (Hasan, 2013). Studies have showed that a group of genes with similar gene expressions are likely to have related gene functions (Mount, 2004). Therefore, how to find the genes that share similar expression patterns across samples is an important question that is frequently asked in the DNA microarray studies (Qin *et al*., 2014).

Clustering, which is a useful technique to constitute unknown groupings of objects (Kaufman and Rousseeuw, 2009), has become an important part of gene expression data analysis (Qin *et al*., 2014; Eisen *et al*., 1998). By investigating the clusters of genes having similar expression patterns across samples, researchers can elucidate gene functions, genetic pathways, and regulatory circuits. Clustering helps to find a distinct pattern for each cluster, as well as more information about functional similarities and gene interactions within the cluster (Hasan and Duan, 2014). For clustering DNA microarray data, a good number of algorithms have been developed that include *k*-means (Tavazoie *et al*., 1999), hierarchical clustering (Eisen *et al*., 1998; Luo *et al*., 2003; Wen *et al*., 1998), self-organizing maps (Tamayo *et al*., 1999; Törönen *et al*., 1999; He *et al*., 2003), support vector machines (Brown *et al*., 2000), Bayesian networks (Friedman *et al*., 2000), and fuzzy logic approach (Woolf and Wang, 2000). In addition to these algorithms, there are others that use genomic information, along with gene expression data, to improve clustering efficiency. Algorithms that fall into this category include an ontology-driven clustering algorithm (Wang *et al*., 2005) and the ones that use information about TS2 upstream regions of the coding sequences and gene expression profiles to get more biologically relevant clusters (Holmes and Bruno, 2000; Barash and Friedman, 2002; Kasturi *et al*., 2003).

Among the existing clustering algorithms, *k*-means and hierarchical clustering algorithms are the most commonly used. *k*-means is computationally faster than hierarchical clustering and produces tighter clusters than the hierarchical clustering algorithm. On the other hand, the hierarchical clustering algorithm computes a complete hierarchy of clusters and hence is more informative than *k*-means. Despite these advantages, 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, and the advantage of hierarchical clustering comes at the cost of low efficiency. Moreover, being computationally expensive, both of these algorithms impede the wide use of these algorithms in gene expression data analysis (Garai and Chaudhuri, 2004; Ushizawa *et al*., 2004; Bolshakova *et al*., 2005). As a solution to this problem, a combined approach was proposed by Chen *et al*. (2005), who first applied the *k*-means algorithm to determine the *k* clusters and then fed these clusters into the hierarchical clustering technique to shorten the merging cluster time and generate a treelike dendrogram. However, this solution still suffers from the limitation of determining the initial value for *k* (Hasan, 2013; Hasan and Duan, 2014).

In this chapter, we propose a new algorithm, hierarchical *k*-means, that combines the advantages of both *k*-means and the hierarchical clustering algorithm to overcome their limitations. Combining different algorithms to overcome their own limitations and produce better results is a popular approach in research (Che *et al*., 2011, 2012; Hasan *et al*., 2012). In this proposed algorithm, initially we applied the hierarchical clustering algorithm and then used the result to decide the initial number of clusters and fed this information into *k*-means clustering to obtain the final clusters. Since similar gene expression profiles indicate similarity in their gene functionalities (Azuaje and Dopazo, 2005), after applying the proposed algorithm to the microarray data set of lung adenocarcinoma using gene ontology (GO) annotations, we explored the change in the enrichment of molecular functionalities of the genes of each cluster for normal tissue and *KRAS*-positive tissues. Our results showed that in each cluster, genes were grouped together based on their expression pattern and molecular functions, which indicate the correctness of this proposed algorithm.

2 METHODS

***k*-means clustering algorithm:** For clustering genes, *k*-means clustering, a well-known method for cluster analysis partition expression levels of *n* genes into *k* clusters, so that the total distance between the cluster's genes and its corresponding centroid, representative of the cluster, is minimized. In short, the goal is to partition the *n* genes into *k* sets S_i, $i=1, 2\ldots, k$ in order to minimize the within-cluster sum of squares (WCSS), defined as

$$WCSS = \sum\nolimits_{j=1}^{k} \sum\nolimits_{i=1}^{n} ||x_i^j - c_j||^2, \tag{4.1}$$

where $||x_i^j - c_j||^2$ provides the distance between a gene and the cluster's centroid.

In this clustering algorithm, the initial cluster centroids are selected randomly. After that, each gene is assigned to the closest cluster centroid. Then each cluster centroid is moved to the mean of the points assigned to it. This algorithm converges when the assignments no longer change. Algorithm 4.1 shows the pseudocode of the *k*-means clustering algorithm.

Hierarchical clustering algorithm: In gene clustering, hierarchical clustering is a method of cluster analysis that builds a hierarchy of clusters (as its name indicates). This clustering method organizes genes into tree structures based on their relation. The basic idea is to assemble a set of genes into a tree, where genes are joined by very short branches if they have very great similarity to each other, and by increasingly long branches as their similarity decreases.

The approaches for hierarchical clustering can be classified into two groups: agglomerative and divisive. The agglomerative approach is a "bottom-up" approach, where each gene starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy. On the other hand, divisive approach is a "top-down" approach, where all genes starts in one cluster and splits are performed recursively as one moves down the hierarchy. In this chapter, we mainly focus on the agglomerative approach for hierarchical clustering.

The first step in hierarchical clustering is to calculate the distance matrix between the genes in the data set. The clustering starts once this matrix of distances is computed. The agglomerative hierarchical clustering technique consists of repeated cycles where the two closest genes having the smallest distance are joined by a node known as a *pseudonode*. The two joined genes are removed from the list of genes being processed and replaced by the pseudonode that represents the new branch. The distances between this pseudonode and all other remaining genes are computed,

ALGORITHM 4.1

k-means

```
Input:   $X = \{x_1,x_2,\ldots,x_n\}$ // set of genes to be clustered.
         $k$ // number of clusters.
Output:  $C = \{c_1,c_2,\ldots,c_k\}$ // set of cluster centroids.
         $L = \{l(x)|x = 1,2,\ldots,n\}$ // set of cluster labels of $X$.

foreach $c_i \in C$ do
    $c_i \leftarrow x_j \in E$ // random selection
end
foreach $x_i \in X$ do
    $l(x_i) \leftarrow calculateMinDistance(x_i, c_j)\ j \in \{1,2,\ldots,k\}$
end
$changed \leftarrow false$
$iter \leftarrow 0$
repeat
    foreach $c_i \in C$ do
        $updateCluster(c_i)$
    end
    foreach $x_i \in X$ do
        $minDist \leftarrow calculateMinDistance(x_i, c_j)\ j \in \{1,2,\ldots,k\}$
        if ($minDist \neq l(x_i)$) then
            $l(x_i) \leftarrow minDist$
            $changed \leftarrow true$
        end
    end
    $iter$++
until ($changed = true$) // no more change in the cluster takes place after the assignment
```

and the process is repeated until only one node remains. Note that there are a variety of ways to compute distances while dealing with a pseudonode: centroid linkage, single linkage, complete linkage, and average linkage. In this chapter, we use average linkage, which defines the distance between two clusters as the average pairwise distance between genes in cluster C_i and C_j calculated using Eq. (4.2):

$$\delta(C_i, C_j) = \frac{\sum_{x \in C_i} \sum_{y \in C_j} \delta(x, y)}{n_i . n_j}, \tag{4.2}$$

where $\delta(x, y)$ is typically given by the Euclidean distance calculated using Eq. (4.3):

$$\delta(x, y) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}. \tag{4.3}$$

The pseudocode of agglomerative hierarchical clustering using average linkage is illustrated in Algorithm 4.2.

ALGORITHM 4.2

Hierarchical Clustering

```
Input:   $G = \{V, E, d\}$ // Weighted graph, $V$ is the set of all genes, $E$ is the set of edge, $d$ is the
weight meaning the distance between two genes.
Output: $T = \{V_T, E_T\}$ // Cluster hierarchy or dendrogram.

$C \leftarrow \{\{v\} \mid v \in V\}$ // Initial clustering. Each gene is placed in separate clusters where each cluster
//contains one gene.
$V_T \leftarrow \{\{v_c\} \mid c \in C\}, E_T \leftarrow \emptyset$ // Initial dendrogram
repeat
    $updateDistanceMatrix(C, G, d)$ //updates the  distance matrix such that distance between the
//new cluster and all remaining clusters are computed.
    $\{C, C'\} \leftarrow$ calculateMinDistance d($C_i$, $C_j$) where $\{C_i, C_j\} \in : C_i \neq C_j$ //calculates the minimum
//distance between cluster $Ci$ and $Cj$
    $C \leftarrow (C \setminus \{C, C'\}) \cup \{C \cup C'\}$ // Merging of clusters to form a new cluster.
    $V_T \leftarrow V_T \cup \{v_{C, C'}\}, E_T \leftarrow E_T \cup \{\{v_{C, C'}, v_C\}, \{v_{C, C'}, v_{C'}\}\}$ //building up the dendrogram
until (|$C$| > 1) // keep doing until only one cluster remains
```

Hierarchical *k*-means: In this proposed algorithm, we selected the value of *k* (i.e., the number of clusters) in a systematic way. Initially, we used the agglomerative hierarchical clustering algorithm for clustering the data set using average linkage and then checked at what level the distance between two consecutive nodes of the hierarchy was the maximum. Using this information, the value of *k* is determined, which is then fed into the *k*-means clustering algorithm to produce the final clusters. In both algorithms, the Pearson correlation coefficient (r) was used as the similarity metric between two samples and $1 - r$ was used as the distance metric. Algorithm 4.3 shows the pseudocode of the proposed algorithm.

ALGORITHM 4.3

Hierarchical *k*-means Clustering

```
Input:   $G = \{V,E,d\}$ // Weighted graph, $V$ is the set of all genes, $E$ is the set of edge, $d$ is the
weight meaning the distance between two genes.
Output: $C = \{c_1,c_2,\ldots,c_k\}$ // set of cluster centroids.
         $R = \{r(v) | v = 1,2,\ldots,n\}$ // set of cluster labels of $V$.

$T \leftarrow hierarchicalClustering(V, E, d)$ // Initial hierarchical clustering that returns a dendrogram $T$
maxDistance $\leftarrow \emptyset$
$l \leftarrow 0$ // at which level the maximum distance with the previous level found in $T$.
$N[] \leftarrow$ nodes in $T$
for $i \leftarrow 2$ to $N.length$
    distance $\leftarrow$ getDistance($N_i$, $N_{i-1}$) //calculating the distance between two consecutive nodes
    if (distance > maxDistance) then
        distance $\leftarrow$ maxDistance
        $l \leftarrow$ level of node $i$ in $T$
end
$k \leftarrow 1 + l$
$(C, R) \leftarrow kMeansClustering(V, k)$
```

3 DATA SET

Lung adenocarcinoma, the most frequent type of non-small-cell lung cancer (NSCLC) accounts for more than 50% of NSCLC, and the percentage is increasing (Okayama *et al*., 2012). Recent studies revealed that activation of the *EGFR, KRAS,* and *ALK* genes defines three different pathways that are responsible for a considerable fraction (30%–60%) of lung adenocarcinomas (Pao and Girard, 2011; Ihle *et al*., 2012; Janku *et al*., 2010; Bronte *et al*., 2010; Gerber and Minna, 2010). The data set used in this research contains expression profiles for 246 samples, of which 20 samples belonged to normal lung tissue. Out of the remaining 226 lung adenocarcinoma samples, 127 were with *EGFR* mutation, 20 with *KRAS* mutation, 11 with *EML4-ALK* fusion, and 68 with triple negative cases. The platform used for this data set was GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. This data set was collected from the GEO database (accession number GSE31210). The data set contained 54,675 genes. In this study, we considered 40 samples consisting of 20 samples from normal tissues and 20 samples from *KRAS*-positive tissues.

To determine the differentially expressed genes, we performed paired student *t*-test (Hsu and Lachenbruch, 2008) and Bonferroni corrections (Bonferroni, 1936), 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 ones, which have adjusted p-values ≤ 0.05. In addition, 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 ≥ 2. Besides this preprocessing, we considered only those genes that are associated with molecular functions according to Gene Ontology (GO).

After performing the t-test, we obtained 21,880 genes having significant p-values ($\leq$ 0.05). We performed Bonferroni correction on these genes and found 1988 genes that had a significantly adjusted p-value ($\leq$ 0.05). Adding the fold change criterion, we reduced the set of differentially expressed genes to 1005. We then performed another step of filtering to keep only those genes that have GO terms and responsible for molecular functions. Finally, we came up with 464 genes in the final data set. The final data set is given partially in Table 4.1, and the complete data set is available in http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31210 (Accessed 06/18/2013).

4 RESULTS AND DISCUSSION

The result of hierarchical clustering for normal tissue data set is shown in Figure 4.1. There are 463 interior nodes in the tree where each node is labeled based on the increasing order of its height. Therefore, the ID for the root is 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 (see Figure 4.2).

Table 4.1 A Brief Overview of the Final data set

		Samples		
Affymatrix ID	Gene Symbol	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

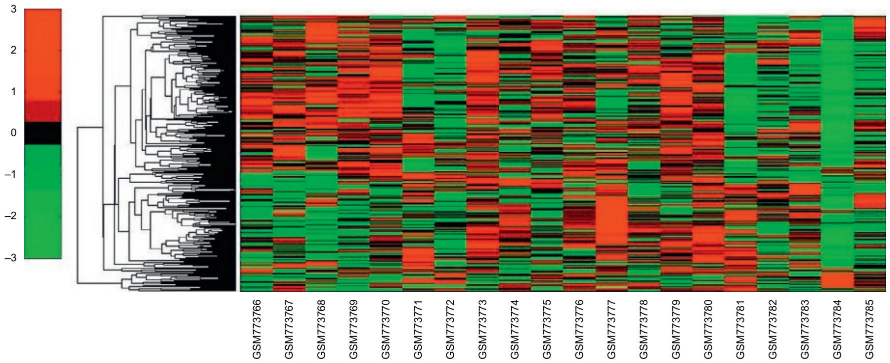

FIGURE 4.1

Hierarchical clustering of the normal tissue data set.

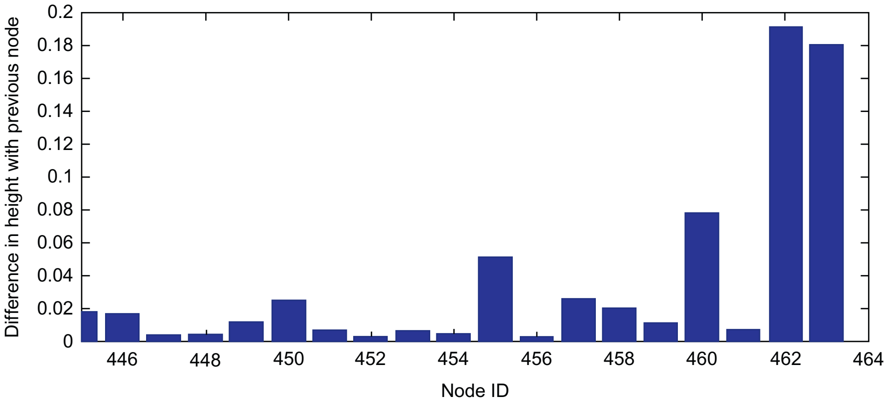

FIGURE 4.2

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.

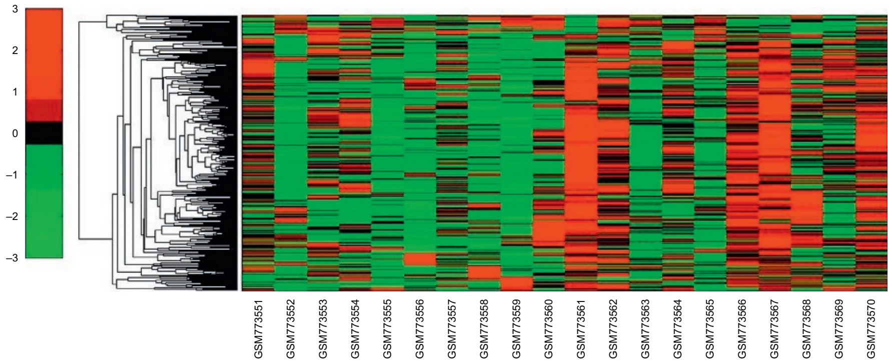

FIGURE 4.3

Hierarchical clustering of the KRAS positive data set.

From Figure 4.2, we can see that the difference is the maximum for nodes 461 and 462. As there is a total of 463 nodes in the tree, node 461 is on level 3 from the top. So, according to the proposed algorithm, the total number of clusters for *k*-means clustering should be 4.

Similarly, the number of clusters for the *KRAS*-positive data set can also be determined. Figure 4.3 shows the hierarchical clustering of *KRAS*-positive data set, and Figure 4.4 shows the height difference between two consecutive nodes. The results indicate that the number of clusters for *KRAS*-positive data set should be 4.

After determining the value for the initial number of clusters (*k*), we passed the value to *k*-means algorithms, and *k* numbers of clusters were formed for both normal and *KRAS*-positive tissues. We explored their common features (genes) and explained the change of molecular function of the genes captured in the clusters

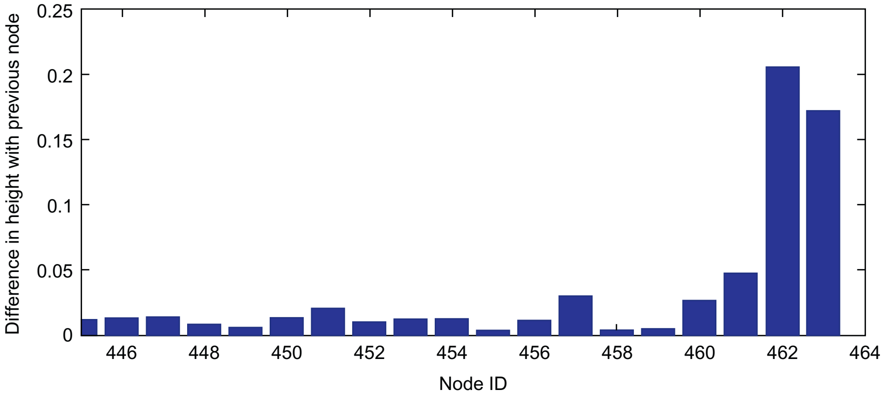

FIGURE 4.4

Height difference between two consecutive nodes in the hierarchical tree generated from a *KRAS*-positive data set.

Table 4.2 List of the Clusters to Be Compared for the Alteration in Molecular Function

Clusters to Compare		
Normal Tissue	***KRAS*-Positive**	**Number of Genes in Common**
Cluster 1	Cluster 1	20
Cluster 2	Cluster 3	52
Cluster 3	Cluster 4	46
Cluster 4	Cluster 2	69

of both normal tissue and *KRAS*-positive tissue using GO annotations. For comparing the molecular function of the clusters of normal tissue and *KRAS*-positive tissues, we took one cluster from the normal tissue data set and one from the *KRAS*-positive data set that have the maximum number of common genes. Table 4.2 shows the clusters that we selected for comparing their molecular functions with the number of genes they have in common.

We explored the molecular functions of the genes in each cluster using GO annotations. Relationships among the genes were represented using a directed acyclic graph (DAG), termed the *GO graph*. We used a web-based tool called the *Gene Ontology Enrichment Analysis Software Toolkit (GOEAST)* (Zheng and Wang, 2008) to generate these graphs. This graph displays enriched Gene Ontology IDs (GOIDs) and their hierarchical relationships in molecular function GO categories. Figures 4.5 and 4.6 show the GO graph for cluster 1 for normal tissue and *KRAS*-positive tissue data set, respectively.

In Figures 4.5 and 4.6, boxes represent GO terms, each labeled by its GOID and term definition. Note that significantly enriched GO terms are shaded yellow. The degree of color saturation of each node is positively correlated with the significance of enrichment of the corresponding GO term. Nonsignificant GO terms within the hierarchical tree are shown as white boxes. In both of these graphs, edges stand for connections between different GO terms. Edges colored in red stand for the relationship between two enriched GO terms, black solid edges stand for the relationship between enriched and unenriched terms, and black dashed edges stand for the relationship between two unenriched GO terms.

In brief, these two figures show that the significant GO terms GO: 0005488 (binding) and GO: 0005515 (protein binding) remain the 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 unenriched in normal tissue, become highly enriched in the *KRAS*-positive tissues, indicating that our proposed algorithm can cluster representative genes of both data sets correctly.

To compare the enrichment status of the two clusters better, we used Multi-GOEAST, which is an advanced version of GOEAST, and it is helpful to identify

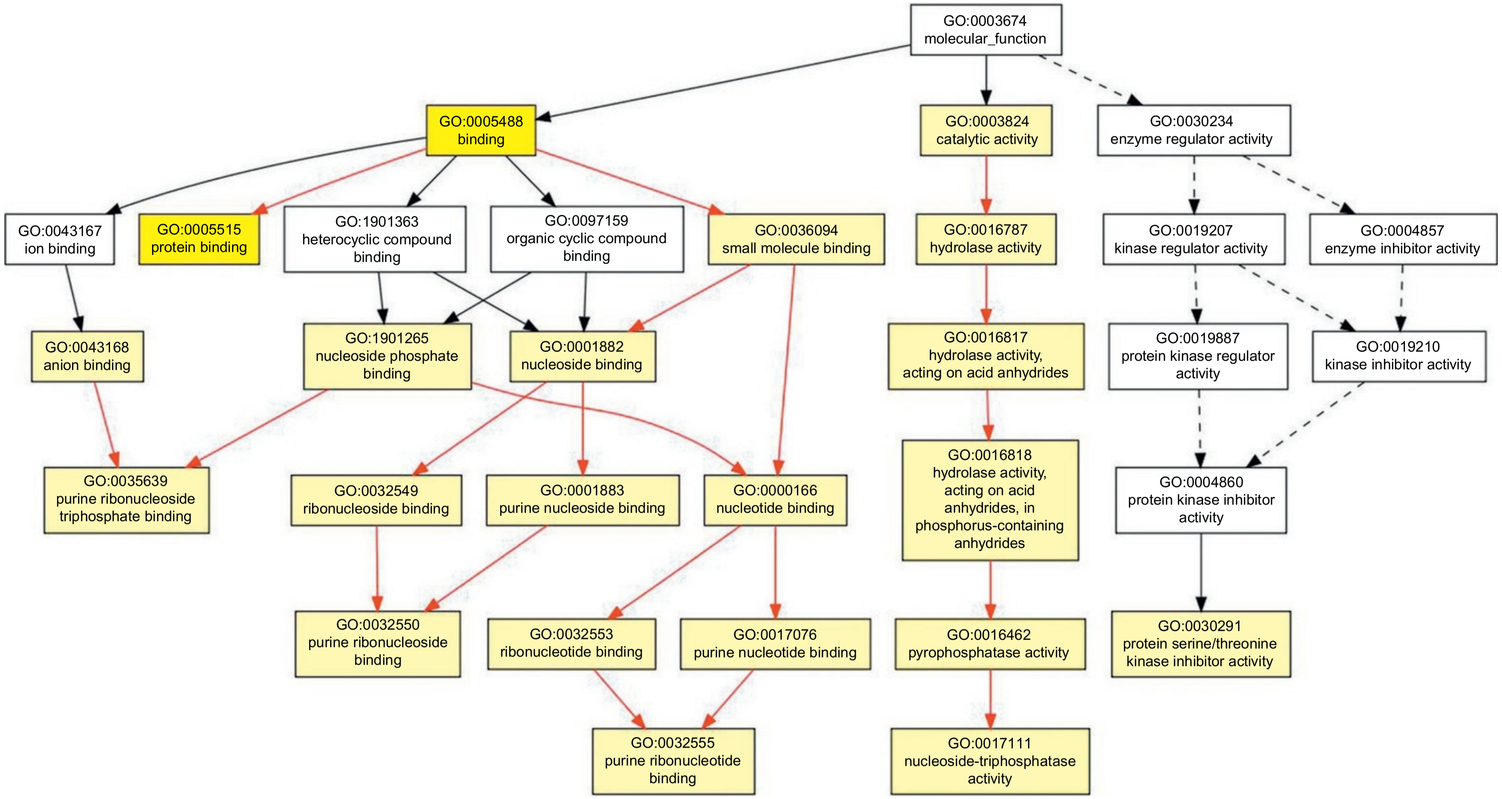
GO:0003674
molecular_function
GO:0005488
binding
GO:0003824
catalytic activity
GO:0030234
enzyme regulator activity
GO:0043167
ion binding
GO:0005515
protein binding
GO:1901363
heterocyclic compound binding
GO:0097159
organic cyclic compound binding
GO:0036094
small molecule binding
GO:0016787
hydrolase activity
GO:0019207
kinase regulator activity
GO:0004857
enzyme inhibitor activity
GO:0043168
anion binding
GO:1901265
nucleoside phosphate binding
GO:0001882
nucleoside binding
GO:0016817
hydrolase activity, acting on acid anhydrides
GO:0019887
protein kinase regulator activity
GO:0019210
kinase inhibitor activity
GO:0035639
purine ribonucleoside triphosphate binding
GO:0032549
ribonucleoside binding
GO:0001883
purine nucleoside binding
GO:0000166
nucleotide binding
GO:0016818
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
GO:0004860
protein kinase inhibitor activity
GO:0032550
purine ribonucleoside binding
GO:0032553
ribonucleotide binding
GO:0017076
purine nucleotide binding
GO:0016462
pyrophosphatase activity
GO:0030291
protein serine/threonine kinase inhibitor activity
GO:0032555
purine ribonucleotide binding
GO:0017111
nucleoside-triphosphatase activity

FIGURE 4.5

GO graph for cluster 1 of the normal tissue data set.

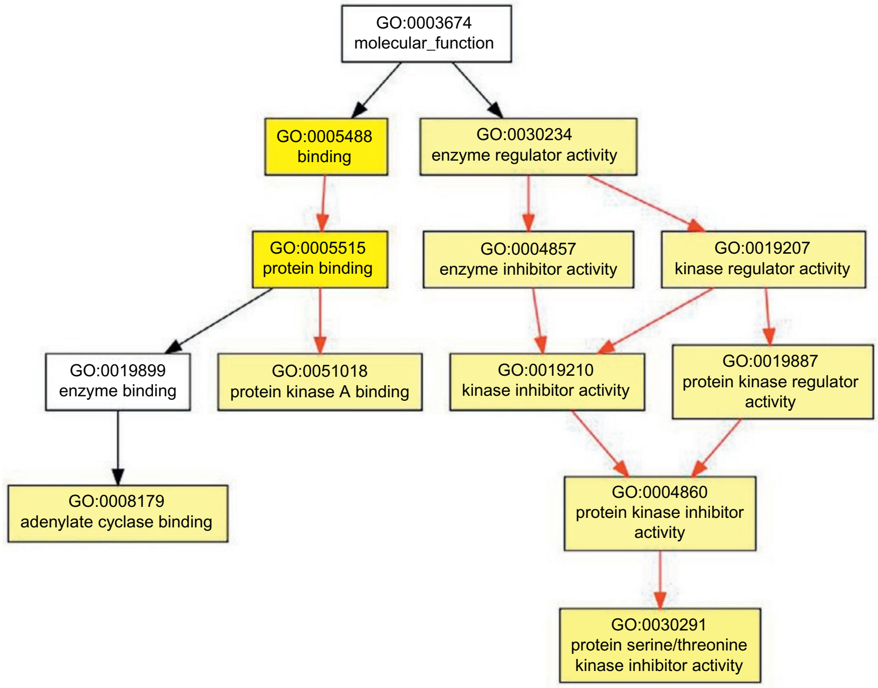
GO:0003674
molecular_function
GO:0005488
binding
GO:0030234
enzyme regulator activity
GO:0005515
protein binding
GO:0004857
enzyme inhibitor activity
GO:0019207
kinase regulator activity
GO:0019899
enzyme binding
GO:0051018
protein kinase A binding
GO:0019210
kinase inhibitor activity
GO:0019887
protein kinase regulator activity
GO:0008179
adenylate cyclase binding
GO:0004860
protein kinase inhibitor activity
GO:0030291
protein serine/threonine kinase inhibitor activity

FIGURE 4.6

GO graph for cluster 1 of the *KRAS*-positive tissue data set.

the hidden correlation between the two clusters (Zheng and Wang, 2008). Figure 4.7 shows the comparative GO graph for cluster 1 of both data sets.

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

Table 4.3 lists the genes associated with the GO terms that are enriched in cluster 1 of the *KRAS*-positive tissue data set, but not in cluster 1 of the normal tissue data set. These GO terms are marked green in the comparative GO graph shown in Figure 4.7. We believe that 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 (see supplementary materials).

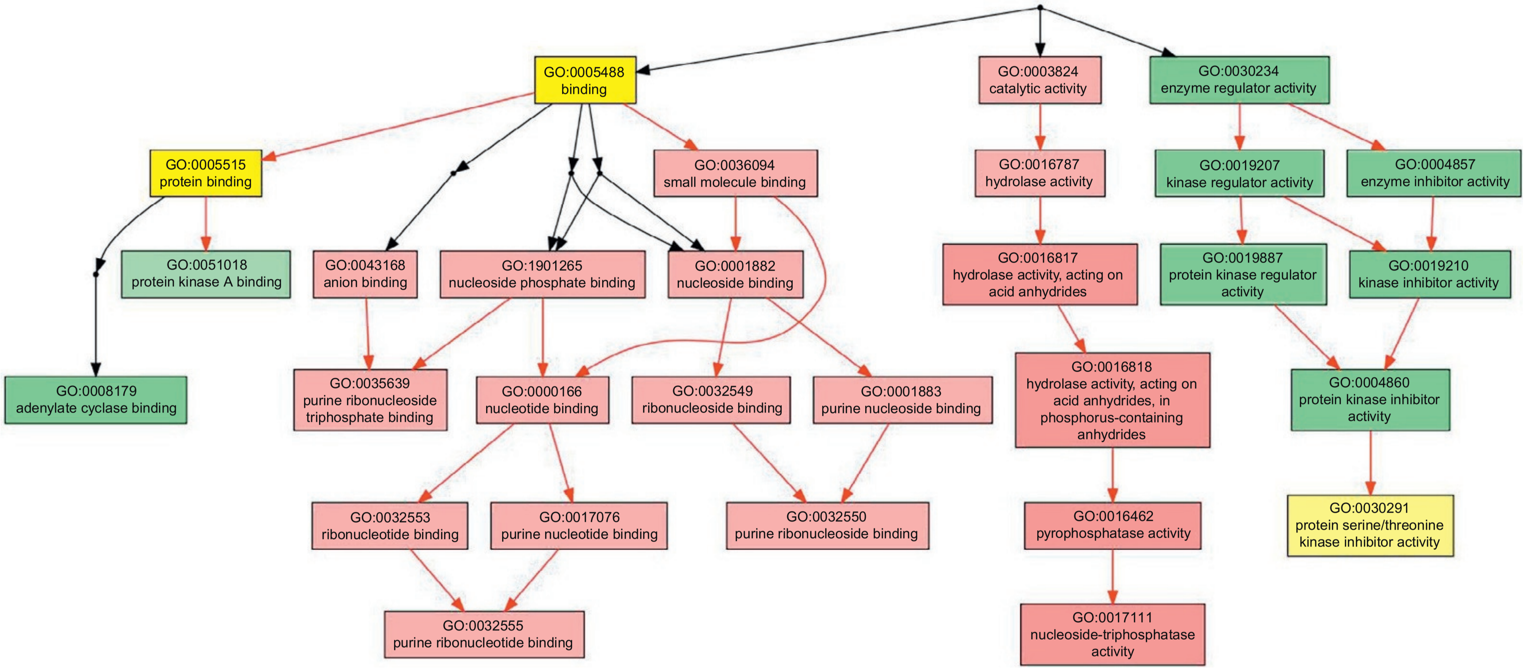
GO:0005488
binding
GO:0003824
catalytic activity
GO:0030234
enzyme regulator activity
GO:0005515
protein binding
GO:0036094
small molecule binding
GO:0016787
hydrolase activity
GO:0019207
kinase regulator activity
GO:0004857
enzyme inhibitor activity
GO:0051018
protein kinase A binding
GO:0043168
anion binding
GO:1901265
nucleoside phosphate binding
GO:0001882
nucleoside binding
GO:0016817
hydrolase activity, acting on acid anhydrides
GO:0019887
protein kinase regulator activity
GO:0019210
kinase inhibitor activity
GO:0008179
adenylate cyclase binding
GO:0035639
purine ribonucleoside triphosphate binding
GO:0000166
nucleotide binding
GO:0032549
ribonucleoside binding
GO:0001883
purine nucleoside binding
GO:0016818
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides
GO:0004860
protein kinase inhibitor activity
GO:0032553
ribonucleotide binding
GO:0017076
purine nucleotide binding
GO:0032550
purine ribonucleoside binding
GO:0016462
pyrophosphatase activity
GO:0030291
protein serine/threonine kinase inhibitor activity
GO:0032555
purine ribonucleotide binding
GO:0017111
nucleoside-triphosphatase activity

FIGURE 4.7

Comparative GO graph for comparing GO enrichment status for cluster 1 of both the normal tissue and *KRAS*-positive data sets.

Table 4.3 GO Terms and Pathways That Are Enriched in Molecular Functions of the Genes of Cluster 1 of *KRAS*-Positive Tissue but Unenriched in the Genes of Cluster 1 of Normal Tissue Data Set

GO ID	GO Term	Associated Genes	Pathway
GO:0030234	Enzyme regulator activity	TIMP3	1 Matrix_metalloproteinases
		CDKN1C	G1_to_S_cell_cycle_reactome
		PAK1	Integrin mediated_cell_adhesion_KEGG
		ECT2	N/A
		RALGPS2	N/A
		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 CONCLUSIONS

In this chapter, we propose hierarchical *k*-means, a new combined clustering algorithm designed to cluster genes in a microarray data set based on their expression levels. In this algorithm, using the output from hierarchical clustering, we systematically determined the value of *k* required for *k*-means clustering. This way, the proposed algorithm overcomes the limitation of *k*-means clustering. This proposed algorithm takes advantage of the ability of hierarchical clustering to get a complete hierarchy of clusters and uses this information in *k*-means clustering to produce tighter clusters.

In this study, we examined 40 samples and 464 genes from the data set of lung adenocarcinoma, which is one of the most frequent types of NSCLC. Out of the 40 samples, 20 were from normal tissue and 20 were from *KRAS*-positive tissue. We applied t-test, Bonferroni correction, and fold change cutoff techniques 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 data set.

After applying the proposed clustering algorithms, we obtained four clusters for both the normal tissue data set and *KRAS*-positive data set. Hereafter, we examined the genes contained in each cluster with respect to their molecular functions based on GO annotation to see what changes in the enrichment of the molecular functions of genes took place from normal tissues to *KRAS*-positive tissues. This way, after checking the change in enrichment of the GO terms, we verified that the proposed algorithm can cluster representative genes of both data sets based on their expression patterns. The coherent approach presented in this chapter shows its correctness to cluster genes, and we believe that it can be generalized for clustering other types of large data sets as well.

REFERENCES

Azuaje, F., Dopazo, J., 2005. Data Analysis and Visualization in Genomics and Proteomics: Wiley Online Library.

Barash, Y., Friedman, N., 2002. Context-specific Bayesian clustering for gene expression data. J. Comput. Biol. 9, 169–191.

Bolshakova, N., Azuaje, F., Cunningham, P., 2005. An integrated tool for microarray data clustering and cluster validity assessment. Bioinformatics 21, 451–455.

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilita. Libreria internazionale Seeber.

Bronte, G., Rizzo, S., La Paglia, L., Adamo, V., Siragusa, S., Ficorella, C., et al., 2010. Driver mutations and differential sensitivity to targeted therapies: a new approach to the treatment of lung adenocarcinoma. Cancer Treat. Rev. 36, S21–S29.

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., et al., 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci. 97, 262–267.

Che, D., Hasan, M.S., Wang, H., Fazekas, J., Huang, J., Liu, Q., 2011. EGID: an ensemble algorithm for improved genomic island detection in genomic sequences. Bioinformation 7, 311.

Che, D., Hasan, M.S., Wang, H., Fazekas, J., Chen, B., Tao, X., 2012. M are better than one: an ensemble method for genomic island prediction. In: International Conference on Bioinformatics and Biomedical Engineering, pp. 426–429.

Chen, T.-S., Tsai, T.-H., Chen, Y.-T., Lin, C.-C., Chen, R.-C., Li, S.-Y., et al., 2005. 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, pp. 405–408.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95, 14863–14868.

Friedman, N., Linial, M., Nachman, I., Pe'er, D., 2000. Using Bayesian networks to analyze expression data. J. Comput. Biol. 7, 601–620.

Garai, G., Chaudhuri, B., 2004. A novel genetic algorithm for automatic clustering. Pattern Recogn. Lett. 25, 173–187.

Gerber, D.E., Minna, J.D., 2010. ALK inhibition for non-small cell lung cancer: from discovery to therapy in record time. Canc. cell 18, 548–551.

Gresham, D., Dunham, M.J., Botstein, D., 2008. Comparing whole genomes using DNA microarrays. Nat. Rev. Genet. 9, 291–302.

Hasan, M.S., 2013. Investigating Gene Relationships in Microarray Expressions: Approaches Using Clustering Algorithms. The University of Akron.

Hasan, M.S., Duan, Z.-H., 2014. A hybrid clustering algorithms and functional study of gene expression in lung adenocarcinoma. In: Proceedings of the World Comp: International Conference on Bioinformatics and Computational Biology, pp. 23–29.

Hasan, M.S., Liu, Q., Wang, H., Fazekas, J., Chen, B., Che, D., 2012. GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences. Bioinformation 8, 203–205.

He, J., Tan, A.-H., Tan, C.-L., 2003. Self-organizing neural networks for efficient clustering of gene expression data. In: Neural Networks, 2003. Proceedings of the International Joint Conference on.pp. 1684–1689.

Holmes, I., Bruno, W.J., 2000. Finding regulatory elements using joint likelihoods for sequence and expression profile data. In: Ismb, pp. 202–210.

Hsu, H., Lachenbruch, P.A., 2008. Paired t test. In: Wiley Encyclopedia of Clinical Trials.

Hunter, L., Holm, L., 1992. Artificial Intelligence and Molecular Biology. AAAI, pp. 866–868.

Ihle, N.T., Byers, L.A., Kim, E.S., Saintigny, P., Lee, J.J., Blumenschein, G.R., et al., 2012. Effect of KRAS oncogene substitutions on protein behavior: implications for signaling and clinical outcome. J. Natl. Canc. Inst. 104, 228–239.

Janku, F., Stewart, D.J., Kurzrock, R., 2010. Targeted therapy in non-small-cell lung cancer—is it becoming a reality? Nat. Rev. Clin. Oncol. 7, 401–414.

Kasturi, J., Acharya, R., Ramanathan, M., 2003. An information theoretic approach for analyzing temporal patterns of gene expression. Bioinformatics 19, 449–458.

Kaufman, L., Rousseeuw, P.J., 2009. Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344 John Wiley & Sons.

Luo, F., Tang, K., Khan, L., 2003. Hierarchical clustering of gene expression data. In: Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on, pp. 328–335.

Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X., Somogyi, R., 1998. Cluster analysis and data visualization of large-scale gene expression data. In: Pacific Symposium on Biocomputing, pp. 42–53.

Mount, D.W., 2004. Sequence and genome analysis. Bioinformatics: Cold Spring Harbour Laboratory Press: Cold Spring Harbour, 2.

Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., et al., 2012. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Canc. Res. 72, 100–111.

Pao, W., Girard, N., 2011. New driver mutations in non-small-cell lung cancer. Lancet Oncol. 12, 175–180.

Qin, L.-X., Breeden, L., Self, S.G., 2014. Finding gene clusters for a replicated time course study. BMC Res. Notes 7, 60.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., et al., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. 96, 2907–2912.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. Nat. Genet. 22, 281–285.

Törönen, P., Kolehmainen, M., Wong, G., Castrén, E., 1999. Analysis of gene expression data using self-organizing maps. FEBS Lett. 451, 142–146.

Ushizawa, K., Herath, C.B., Kaneyama, K., Shiojima, S., Hirasawa, A., Takahashi, T., et al., 2004. cDNA microarray analysis of bovine embryo gene expression profiles during the pre-implantation period. Reprod. Biol. Endocrinol. 2, 77.

Wang, H., Azuaje, F., Bodenreider, O., 2005. An ontology-driven clustering method for supporting gene expression analysis. In: Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on, pp. 389–394.

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10, 57–63.

Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., et al., 1998. Large-scale temporal gene expression mapping of central nervous system development. Proc. Natl. Acad. Sci. 95, 334–339.

Woolf, P.J., Wang, Y., 2000. A fuzzy logic approach to analyzing gene expression data. Physiol. Genom. 3, 9–15.

Zheng, Q., Wang, X.-J., 2008. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. 36, W358–W363.

SUPPLEMENTARY MATERIALS

List of genes of the clusters and detailed results from GO enrichment analysis are presented in the supplementary figures which can be found at http://www.cs.uakron.edu/~duan/Chapter04/SupplementaryMaterials.pdf.