Journal of Data Science **2**(2004), 297-309

Identifying the Patterns of Hematopoietic Stem Cells Gene Expressions Using Clustering Methods: Comparison and Summary

Jie Chen[1], Xi He[2], and Linheng Li[2]
[1] *University of Missouri and* [2] *Stowers Institute for Medical Research*

Abstract: Clustering algorithms have been used to analyze microarray gene expression data in many recent applications. In this paper, we make a comparison among popularly used clustering methods, including hierarchical clustering with average, complete, and single linkages, k-means clustering, k-means clustering with hierarchical initialization, and self organization map (SOM), by making use of our hemotopietic stem cell (HSC) microarray data. To understand the biological pathways from HSC to proliferative multipotent progenitor (MPP), and from MPP to either common lymphoid progenitor (CLP) or common myeloid progenitor (CMP), statistical clustering is an important tool. Our results demonstrated that the HSC microarray data set casts some challenge on clustering algorithms as different clustering algorithms resulted in clusters that were not all consistent. We compared the results by using the total within-cluster sum of squares of dispersions and the biological functions of the genes, and reached the conclusion that k-means clustering with hierarchical average or complete linkage initialization performed the best among all the methods we compared. Our investigation of the clustering methods with HSC microarray data provide a useful approach and guide to medical researchers who use clustering algorithms in analyzing their microarray or related data sets.

Key words: Clustering algorithms, Hematopoietic stem cells, hierarchical clustering methods, k-means clustering, microarray, self-organization map.

1. Introduction

Microarray technology has made it possible to quantify the expression patterns of thousands or tens of thousands of genes in various tissues, cell lines, and conditions simultaneously. When such abundant numerical information becomes available, statistical data analysis is inevitably needed to help biological scientists to organize, summarize, and interpret the results inherent in the numerical expressions of genes. One frequently used method, is the statistical clustering method/algorithm in microarray data analysis. A seminal paper by Eisen *et al.*

(1998) used average-linkage hierarchical clustering to identify patterns in a budding yeast *Saccharomyces Cerevisiae* microarray data set. They found that genes with similar biological functions cluster together. Several other authors, such as, Kerr and Churchill (2001), Michaels *et al.* (1998), also used clustering to analyze different microarray data sets. Self-organization map (SOM), a statistical data mining tool, has also been used in microarray data analysis such as in Tamayo *et al.* (1999). In the recent paper of Akashi *et al.* (2003), the k-means clustering with pre-determined initial seeds (genes) of known biological functions was used for clustering HSC microarray data. Biological pathways involved with many important genes were revealed clearly through the clusters.

The expression patterns of genes contain valuable information about the genes; how to retrieve this information becomes a challenge to statisticians/biostatisticians. Because of the abundance of information inherent in the gene expressions, and because of the complexity of the gene regulation networks and pathways, clustering is an important statistical tool for organizing the information and condensing the genes into smaller groups/clusters with similar expression patterns so that further biological study of the genes can be conducted.

In the following we first introduce our hematopietic stem cell (HSC) microarray data, then we discuss and review several popularly used clustering algorithms in microarray data analysis and employ the total within-cluster sum of squares of dispersions (TWSS) as an evaluation criterion for comparing different clustering methods. Then, using our HSC data we present the TWSS values for the clusters obtained by different algorithms. Finally, we discuss how to choose a suitable clustering algorithm for a microarray data set.

2. Hematopoietic Stem Cell (HSC) Microarray Data

Hematopoietic stem cells (HSCs) are clonogenic cells, which possess the properties of both self-renewal and multilineage potential for giving rise to all types of mature blood cells. Early HSC development displays a hierarchical arrangement starting with HSCs, which have extensive self-renewal capability. Next is the expansion stage, corresponding to proliferative multipotent progenitor (MPP). MPP is also a stage of priming or preparation for differentiation. MPP then commits to either common lymphoid progenitor cells (CLP), which give rise to all the lymphoid lineages, or common myeloid progenitor cells (CMP), which produce all the myeloid lineages. We obtained stem-cell samples from mice and used Affymetrix Microarray Chips to read the expressions of the genes in HSCs, MPPs, CLPs, and CMPs. For more details of the experiment and data production, please refer to Akashi *et al.* (2003). As understanding the transcriptional accessibility for multi-tissue and multi-hematopoietic lineage genes is our primary goal, we were interested in how the genes were clustered together and how they

were associated with each other biologically.

3. Clustering Methods in Analyzing the Stem-Cell Gene Expression Data

Clustering is a primitive multivariate technique in that no assumptions are made concerning the number of groups or the group structure. The "natural" grouping (clustering) depends on the definition of similarity. All possible groupings are not feasible as the number $N(n, c)$ of different partitions of n objects into c clusters is the Sterling number of the second kind (Liu, 1968), namely, $N(n, c) = O(n^t)$ for some $t > 1$. This number $N(n, c)$ becomes quite large even for moderate n and c. As pointed out in Everitt *et al.* (2001), even with today's advanced computer technology, it remains impractical to enumerate all possible clusters. Therefore, there is some subtlety left when clustering is applied to any data set especially for such large data set as microarray data set. We begin our general review and description of the clustering methods in the sequel below. For a more detailed discussion on clustering, the reader is referred to the monograph by Everitt *et al.* (2001).

3.1 Hierarchical clustering algorithms

Hierarchical clustering has two approaches: Agglomerative method and divisive method. Agglomerative method starts with n clusters for a data set of n observations, and then joins two "nearest" observations into one cluster in the next step according to a selected similarity measure. This process continues until all the observations merge into one cluster. The divisive method starts with one cluster, then the two objects that are most "dissimilar" are separated to form two clusters; this process continues until all the n observations form n clusters (one observation per cluster). The similarity/dissimilarity measures used in hierarchical clustering are usually one of the following distance measures: Euclidean distance, squared Euclidean distance, Pearson's correlation coefficient, city block distance (Manhattan distance), or Minkowski distance. Hierarchical clustering procedures are associated with different linkage methods. Three popularly used linkage methods are: single linkage (minimum distance or nearest neighbor), complete linkage (maximum distance or farthest neighbor), and average linkage (average distance). Another method is Ward's method (see Lattin *et al.* (2003) for details), which is based on minimizing the "loss of information" from joining two clusters. In addition, the centroid method is used sometimes.

The hierarchical method is now popularly used in analyzing microarray data after the publication of the paper by Eisen *et al.* (1998). We would like to point out that the hierarchical clustering method is highly structured and the

clustering results depend on which linkage method and distance measure is used. This problem was recently discussed and illustrated through gene expression data by Goldstein *et al.* (2002). As pointed out in Chen *et al.* (2002), the average linkage hierarchical method was the worst among the four clustering methods they compared. According to Chen *et al.* (2002), k-means clustering outperformed average linkage clustering most of the time in most cases of their embryonic stem (ES) cell data. This conclusion is in line with what we observed when we analyzed our HSC microarray data in Akashi *et al.* (2003). In the following section, we make a summary about the k-means clustering method.

3.2 The k-means clustering

Another approach of clustering, rather than the hierarchical clustering, is the partitioning method such as k-means clustering, and the partitioning around medoids (PAM) method (Chen *et al.*, 2002). Let $\mathbf{X}$ be an $n \times p$ data matrix, with each row a $1 \times p$ observation vector. The total dispersion matrix due to partitioning n observation vectors into c clusters can be written as

$$\mathbf{T} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})',$$

where $\mathbf{x}_{ij}$ is the p-dimensional observation vector of the jth object in cluster i, $\overline{\mathbf{x}}$ is the p dimensional vector of overall sample means for each variable, and n_i is the size of cluster i. $\mathbf{T}$ can be partitioned into $\mathbf{T} = \mathbf{W} + \mathbf{B}$, where $\mathbf{W}$ is the within-cluster dispersion matrix and $\mathbf{B}$ the between-cluster dispersion matrix. Let $\overline{\mathbf{x}}_i$ be the p-dimensional vector of sample means within cluster i, then:

$$\mathbf{W} = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)(\mathbf{x}_{ij} - \overline{\mathbf{x}}_i)',$$

and

$$\mathbf{B} = \sum_{i=1}^{c} n_i (\mathbf{x}_{ij} - \overline{\mathbf{x}})(\mathbf{x}_{ij} - \overline{\mathbf{x}})'.$$

The purpose of clustering is to group objects that are homogenous (using selected criterion) into one cluster and to separate objects that are heterogeneous into different clusters. Many clustering methods employ the criterion of either maximizing the between-cluster dispersion or minimizing the within-cluster dispersion. This approach results in three criteria in terms of the matrices $\mathbf{W}$ and $\mathbf{B}$, namely: minimization of trace($\mathbf{W}$) (or maximizing trace($\mathbf{B}$)) as discussed in MacQueen (1967), and Ball and Hall (1967); minimization of det($\mathbf{W}$) (equivalently, maximization of det($\mathbf{T}$)/det($\mathbf{W}$)) as in Friedman and Rubin (1967), and

Marriott (1971, 1982); maximization of trace($\mathbf{BW}^{-1}$) as in Friedman and Rubin (1967).

The k-means algorithm is one such algorithm that minimizes trace($\mathbf{W}$). The k-means clustering algorithm has its different versions. MacQueen (1967) proposed an algorithm that assigns items to the cluster having the nearest centroid according to the Euclidean distance. It starts with an initial partition of k clusters, or k initial centroids (seed points). It proceeds through the list of items, assigning an item to the cluster whose centroid is nearest. It then recalculates the centroid for the cluster receiving the new item and for the cluster losing the item. The process is repeated until no more reassignment of items occurs. This algorithm has been implemented by a variety of commercial statistical software, such as Minitab, that can perform k-means clustering.

3.3 Self-organization map (SOM)

SOM is popularly used in statistical data mining. It starts with a predetermined geometry of the clusters and then maps the observations to its nearest centroid. The purpose of SOM is to geometrically map the high dimensional input data into a good lower dimensional grid of nodes. There is a detailed illustration of SOM and free software available in Tamayo *et al* (1999). Affymetrix also provides SOM for people who use the Affymetrix Data Mining tool to organize the data obtained from Affymetrix microchips. We will give a comparison between clustering and SOM in a later section of this paper.

3.4 The total within-cluster sum of squares of dispersions (TWSS)

Clustering methods are exploratory tools for data analysis, and there is no formal statistical inference in any of the methods. When the data set is not large, all the clustering methods seem to produce consistent results after applying several runs of any clustering method mentioned above. The challenge emerges with the information era when huge (several thousands or several hundreds of thousands) data sets such as microarray data sets are readily available. Chen *et al.* (2002) suggested a couple of indices to evaluate some clustering methods using embryonic stem (ES) cell data. We took into consideration the total sample variance approach as used in traditional statistical inferences including regression, ANOVA, and MANOVA. It is a natural measure of the variation arouse in any data set.

As the total variance of matrix $\mathbf{T}$, trace($\mathbf{T}$), is a fixed constant, we propose to use the total variance of the sum of squares of the within cluster dispersion matrix $\mathbf{W}$ as an evaluation index, denoted by TWSS. That is $TWSS$ =trace($\mathbf{W}$). According to the nature of the TWSS index, a clustering method with smaller

TWSS value is a better algorithm as a smaller TWSS value indicates better homogeneity in each cluster. This index will be used in the following data analyses.

4. Applying Clustering Algorithms to the HSC Microarray Data

4.1 HSC microarray data pre-screening

The expressions of 12,429 genes in HSCs, MPPs, CLPs, and CPMs were obtained by using Affymetrix chip A. As any clustering method is sensitive to outliers or "noise points" in a data set, we deleted some outlier genes whose expression levels are above 20000. For illustration purposes, we selected genes that had passed the following filter as differentially expressed genes and then the clustering method was applied to these differentially expressed genes. The genes in our microarray data were considered as differentially expressed if they passed the filter given by

$$\left|y_{j(m)} - y_{j(l)}\right| > 100 \text{ and } y_{j(m)}/y_{j(l)} > 4, \text{ for } j = 1, \ldots, n,$$

where $y_{j(m)}$ and $y_{j(l)}$ are the order statistics with $y_{j(1)} \leq \ldots \leq y_{j(m)}$for the jth gene. This filtering criterion simultaneously considered the absolute difference of the gene expression levels and the fold change of the expression levels for each gene across the conditions. There were 659 genes that passed this filter. We also normalized these 659 genes such that each gene had a mean of 0 and a standard deviation of 1.

4.2 Comparison among the clustering methods in light of the HSC data

The k-means clustering result depends on the initial partition (initial seeds) of the clusters (Milligan, 1980). However, if the initial seeds are selected according to some known features of the data, then the k-means clustering is quite robust according to Milligan (1980). This point was addressed in Akashi *et al.* (2003).

When the initial seeds of clusters are not available, we design to use one of the hierarchical algorithms to generate seeds for different clusters. Once the seeds are obtained, one can proceed to the k-means clustering method to achieve final partition of the gene expression levels. This design/approach of clustering makes use of the merits of hierarchical algorithms and k-means method and overcomes their shortfalls simultaneously. This approach is illustrated in the following analysis of the 659 genes observed in HSC, MPP, CLM, and CMP in Chip A.

We applied the average linkage hierarchical algorithm, k-means algorithm with average linkage hierarchical initialization, complete linkage hierarchical algorithm, k-means algorithm with complete linkage hierarchical initialization, sin-

gle linkage hierarchical algorithm, and k-means algorithm with single linkage hierarchical initialization, for different k values: $k = 15, 18, 21, 24$, and 27. The distance measure we used is the squared Euclidean distance, which is equivalent to Pearson's correlation as we have normalized each gene to have mean 0 and variance 1. All the calculations were performed using Minitab Data Analysis software. Finally, we also used SOM to cluster these 659 genes by employing $5 \times 3, 3 \times 5, 6 \times 3, 3 \times 6, 7 \times 3, 3 \times 7, 8 \times 3, 3 \times 8, 9 \times 3$ and 3×9 as the starting geometry. The SOM results were obtained by using the free software provided in Tamayo *et al.* (1999).

The TWSS values were calculated for all the clustering methods used. These values are graphed in Figure 1 (A-F). Figure 1A shows that the k-means clustering with average linkage hierarchical initialization is better than the corresponding average linkage hierarchical clustering for any cluster size. Figure 1B shows that k-means clustering with complete linkage hierarchical initialization is better than the corresponding complete linkage hierarchical clustering for any cluster size. Figure 1C shows that k-means clustering with single linkage hierarchical initialization is better than the corresponding single linkage hierarchical clustering for any cluster size. It is important to know that single linkage hierarchical clustering is the worst (with largest TWSS in all cases), and k-means clustering can even correct the situation brought up by single linkage (see Figure 1C). Figure 1D offers a comparison between SOM with starting geometry of $3\times5, 3\times6, 3\times7, 3\times8, 3\times9$, and SOM with starting geometry of $5\times3, 6\times3, 7\times3, 8\times3, 9\times3$. When the cluster size gets larger, the former starting geometry tends to give better clustering results. In Figure 1E, we compared all six clustering methods (including a k-means clustering with random initialization) except the single linkage method, which is the worst one as we noted before. In Figure 1E, we observe that even k-means with random initialization performs well in terms of TWSS. As the cluster size gets larger, k-means clustering with complete linkage hierarchical initialization and k-means clustering with average linkage hierarchical initialization performed almost identically the best. Finally, in Figure 1F, we compared the SOMs with all six methods mentioned above. It is very clear that the six methods are better than the SOMs in general.

Now that we have some *ad hoc* comparisons of the clustering methods, we need to know how the biological information is conveyed by these clustering methods. This is explored in the following section.

4.3 Biological interpretation of the clustering results

We were particularly interested in four patterns that would appear in some of the clusters: 1. Genes predominantly expressed in HSCs which are involved

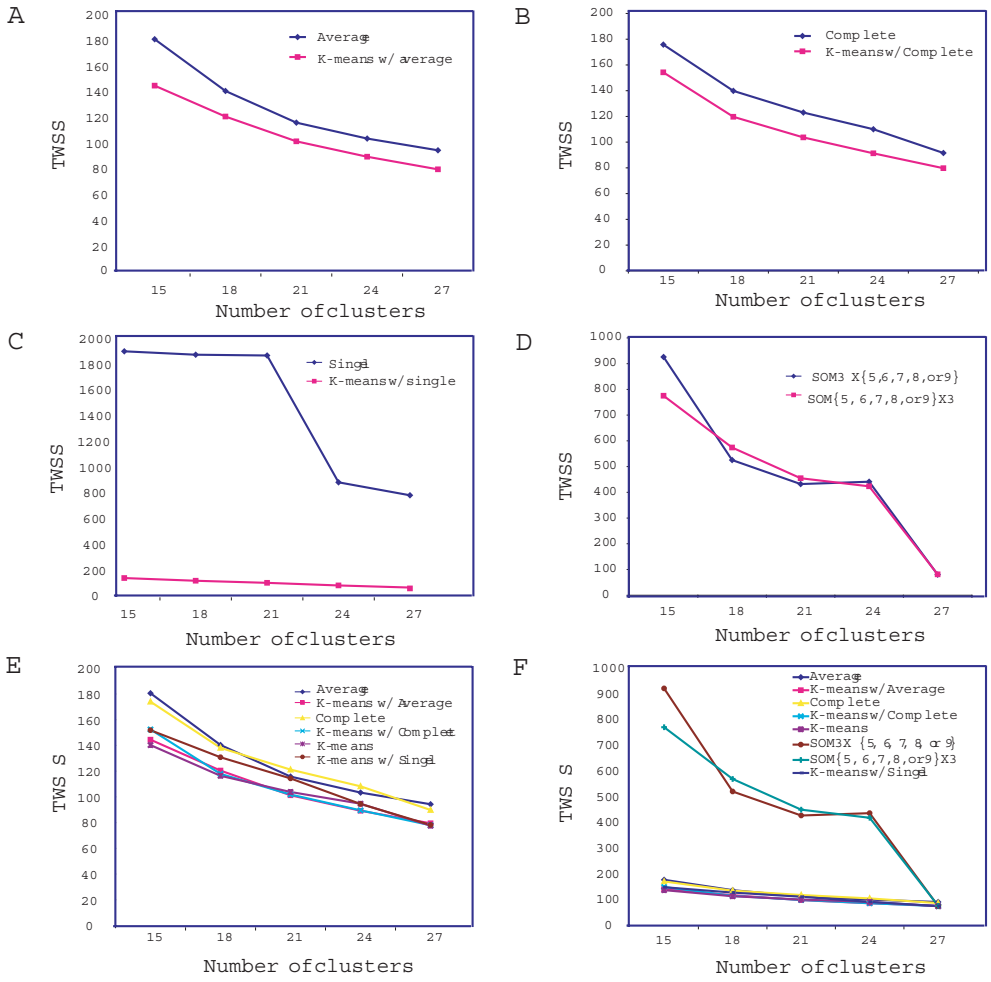

Figure 1. TWSS values for the clusters generated by different clustering methods at different cluster sizes are graphed from A to F for the indicated comparisons shown in each graph. The smaller the TWSS value is, the better the cluster is.

in maintaining stem cell compartment; 2. Genes that are preferentially upregulated in MPPs because MPPs are highly proliferative cells and presumably at a priming stage for both myeloid and lymphoid differentiation; 3. Genes that are upregulated in CLP; and, 4. Genes that are upregulated in CMPs. Due to space limitations and for simplicity and illustration purposes, we only show comparison of the clustering results ($k = 27$) obtained by complete linkage hierarchical clustering and k-means clustering initialized by complete linkage.

There are 123 genes predominantly expressed in HSCs. They are identified in Clusters 4, 6, and 7 by complete linkage hierarchical and k-means with complete

Table 1: Predominantly expressed genes in HSC identified by k-means with complete linkage initialization only

gene	description
AB008516	Tetratricopeptide repeat domain
AF006688	Muspaox
AI154017	Mus musculus cDNA
AI849587	Mus musculus cDNA
AJ005563	Small proline-rich protein 2E
AV007820	Mus musculus cDNA
AV235418	Mus musculus cDNA
AW121616	Mus musculus cDNA
AW124933	Mus musculus cDNA
U16175	Thrombospondin 3
U58972	Growth factor independent 1
U72644	Lymphocyte specific transcript (LST)
U90435	Flotillin
Y14004	Acyl-CoA thioesterase
Z12604	Matrix metalloproteininase 11
AA163908	Mus musculus cDNA
AF011336	Putative E1-E2 ATPase
AI840198	Mus musculus cDNA
M32032	Selenium binding protein 1
U73478	Acidic nuclear phosphoprotein 32
U83509	angiopoietin-1
X12761	Jun oncogene
X61597	Kallikrein-binding protein
M90388	Tyrosine phosphatase
U16985	Lymphotoxin-beta

linkage. These genes are given in Figure 2A. Some of these genes may play a role in maintaining stem cells, such as transcriptional factors LRG-21, LKLF, TCF-3, and growth factors Wnt1 and Dhh (see details in Akashi *et al.*, 2003). In addition, k-means with complete linkage identified 26 more genes predominantly expressed in HSC than complete linkage clustering alone. These genes are listed in Table 1. Further, k-means with complete linkage also relocated 6 genes, which are identified by complete linkage hierarchical as in this category, into other category.

There are 47 genes that are preferentially upregulated in MPPs. They are identified in Clusters 11, 14, and 21 by both complete linkage hierarchical and k-means with complete linkage. These genes are given in Figure 2B. Some of

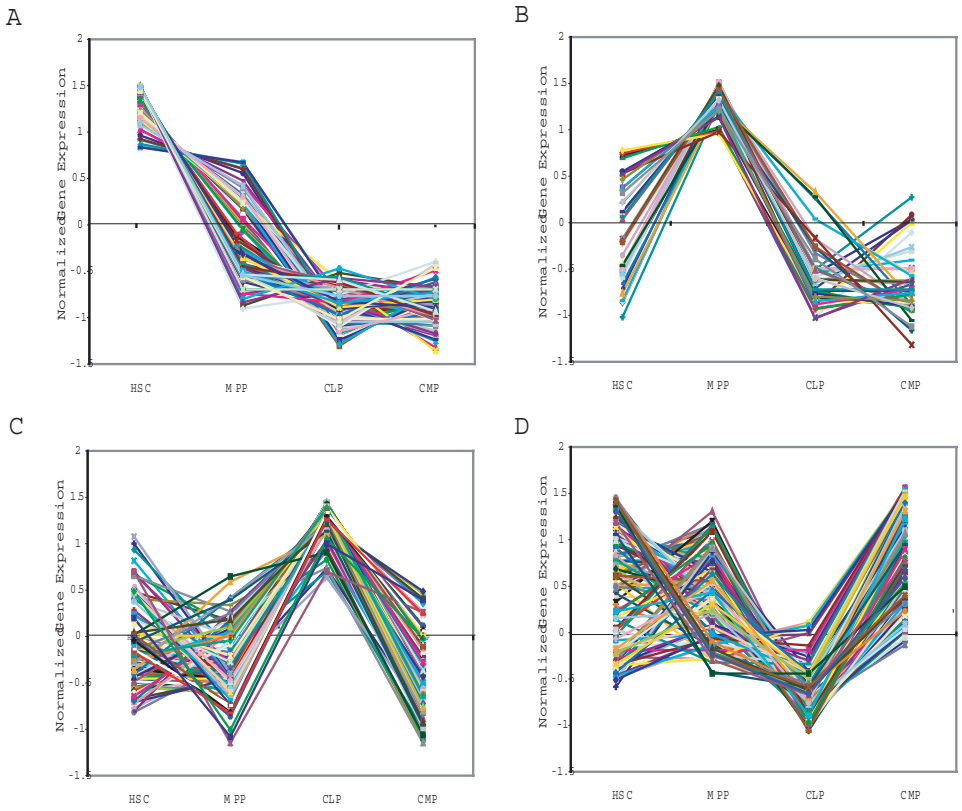

Figure 2. The normalized gene expressions are graphed from A to D. The pathways of the genes are shown in each graph.

these listed genes make biological sense. Leukotriene B4 receptor and Cyclin E are involved in the regulation of cell proliferation. Differentiation factors, such as myeloid associated differentiation protein, erythroid differentiation regulator, and growth differentiation factor 9, are required for cell differentiation. The k-means clustering with complete linkage also identified two more novel genes that are preferentially upregulated in MPPs (AW061073 and AW227620). The k-means with complete linkage also relocated 1 gene, which was also identified by complete linkage hierarchical, into another category.

There are 146 genes that are upregulated in CLP. They are identified in Clusters 2, 10, 12, 20, 23, 25, and 26 by both complete linkage hierarchical and k-means with complete linkage. These genes are given in Figure 2C. This cluster is enriched with genes that are essential for the commitment of lymphoid lineages or have important immunological functions. Examples of these genes are: IL-12a, IL-18R, TCR, CD94, CD7, CD28, Rag-1, Notch1 and Hey1 (see

detailed description in Akashi *et al.*, 2003). The k-means with complete linkage also identified two more genes that are upregulated in CLP: NAD-dependent methylenetetrahydrofolate dehydrogenase and AI842887.

There are 141 genes that are upregulated in CMP. They are identified in Clusters 3, 8, 13, 19, 22, 24, and 27 by both complete linkage hierarchical and k-means with complete linkage. These genes are given in Figure 2D. The functions of most of these listed genes, including transcriptional factors C/EBPdelta, NF-E2, Lim only 2, cytokine receptors G-CSFR, IL-11R, functional proteins myeloperoxidase and Plysozyme, are associated with myeloid differentiation or function (see details in Akashi *et al.*, 2003). Further, k-means with complete linkage also identified seven more genes that are upregulated in CMP and relocated 25 genes, which were identified by complete linkage hierarchical, into another category.

5. Conclusion and Discussion

Clustering methods are useful in recognizing gene expression patterns in large gene profiling. Careful selection of clustering methods in analyzing microarray data is important. Different hierarchical clustering methods do not always give the same clustering results for one data set (Johnson and Wichern, 1998) because of the similarity/dissimilarity and distance functions employed in these methods. This problem may become severe if the data set is huge (as is the case of a microarray data set). Hierarchical methods tend to take a long time for computation and some produce inversions. An inversion occurs when an object joins an existing cluster at a smaller distance (greater similarity) than that of a previous consolidation. Once two observations are joined in a cluster by hierarchical clustering, they can never be separated, while in k-means clustering they still can be separated if they are closer to some new centroids of new clusters. The k-means clustering method is not only an optimal clustering algorithm, but is also much faster than the hierarchical methods in computation for large data sets. Furthermore, the k-means clustering gives more compact clusters than hierarchical methods. When initial seeds of clusters are chosen to be genes of vital and known biological functions, the k-means clustering is robust and produces meaningful results. Alternatively, when seed genes are not available or hard to find, the k-means clustering with initialization by either complete or average linkage hierarchical clustering performs the best among other clustering methods as well.

Acknowledgment

We thank Dr. Ke Xia for his enormous assistance on database analysis, and D. di Natale for her assistance on manuscript proofreading.

References

Akashi, K., He, X., Chen, J., Iwasaki, H., Niu, C., Steenhard, B., Zhang, J., Perera, R., Haug, J., and Li, L. (2003). Transcriptional Accessibility for Multi-Tissue and Multi-Hematopoietic Lineage Genes is Hierarchically Controlled During Early Hematopoiesis. *Blood* **101**, 383-390.

Ball, G. H. and Hall, D. J. (1967). A Clustering Technique for Summarizing Multivariate Data. *Behavioural Science* **12**, 153-155.

Chen, G, Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S. H., and Zhang, M. Q. (2002). Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data. *Statistica Sinica* **12** , 241-262.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863-14868.

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster Analysis.* 4th ed., Arnold, London.

Friedman, J.H. and Rubin, J. (1967). On Some Invariant Criteria for Grouping Data. *Journal of the American Statistical Association* **62**, 1159-1178.

Goldstein, D. R., Ghosh, D., and Conlon, E. M. (2002). Statistical Issues in the Clustering of Gene Expression Data. *Statistica Sinica* **12** , 219-240.

Johnson, R. A. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis.* Prentice Hall.

Kerr, M. K. and Churchill, G. A. (2001). Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments. *Proceedings of the National Academy of Sciences* **98**, 8961-8965.

Lattin, J., Carroll, J. D., and Green, P. E. (2003). *Analyzing Multivariate Data.* Thomson-Brooks/Cole.

Liu, C. L. (1968). *Introduction to Combinatorial Mathematics.* McGraw-Hill, New York.

MacQueen, J. B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability.* 281-297.

Marriott, F. H. C. (1971). Practical Problems in a Method of Cluster Analysis. *Biometrics* **27**, 501-514.

Marriott, F. H. C. (1982). Optimization Methods of Cluster Analysis. *Biometrika* **69**, 417-421.

Michaels, G. M., Carr, D. B., Askenazi, M., Fuhrman, S., Wen, X., and Somogyi, R. (1998). Cluster Analysis and Data Visualization of Large-Scale Gene Expression Data. *Pacific Symposium on Bicomputing* **3**, 42-53.

Milligan, G. W. (1980). An Examination of the Effects of Six Types of Error Perturbation on Fifteen Clustering Algorithms. *Psychometrika* **45**, 325-342.

Tamayo, P., Slonim, A., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences* **96**, 2907-2912.

Received March 11, 2003; accepted September 3, 2003.

Jie Chen
Department of Mathematics and Statistics
University of Missouri – Kansas City
Kansas City, MO 64110, USA
chengj@umkc.edu

Xi He
Stowers Institute for Medical Research
Kansas City, MO 64110, USA

Linheng Li
Stowers Institute for Medical Research
Kansas City, MO 64110, USA