

Cluster stability scores for microarray data in cancer studies

Mark Smolkin and Debashis Ghosh

Department of Biostatistics, University of Michigan

Corresponding author:

Debashis Ghosh, Ph.D.

Department of Biostatistics

School of Public Health, University of Michigan

1420 Washington Heights, Room M4057

Ann Arbor, Michigan 48109-2029

Phone: (734) 615-9824

Fax: (734) 763-2215

Email: ghoshd@umich.edu

Abstract

Motivation: A potential benefit of profiling of tissue samples using microarrays is the generation of molecular fingerprints that will define subtypes of disease. Hierarchical clustering has been the primary analytical tool used to define disease subtypes from microarray experiments in cancer settings. Assessing cluster reliability poses a major complication in analyzing output from clustering procedures. We address this problem by developing cluster stability scores using subsampling techniques. These scores exploit the redundancy in biologically discriminatory information on the chip. Our approach is generic and can be used with any clustering method. We propose procedures for calculating cluster stability scores for situations involving both known and unknown numbers of clusters.

Results: The method is illustrated by application to data three cancer studies; one involving malignant melanoma (Bittner et al., 2000), the second involving B-cell lymphoma (Alizadeh et al., 2000), and the final is from a childhood cancer study (Khan et al., 2001).

Availability: Code implementing the proposed analytic method can be obtained at the second author's website.

Contact: ghoshd@umich.edu

Introduction

Due to the advent of high-throughput microarray technology, scientists have been able to conduct global molecular profiling studies. One of major disease areas in which microarrays have been utilized has been in cancer (Alizadeh et al., 2000; Bittner et al., 2000; Khan et al., 2001). One of the scientific goals of these experiments is the discovery of disease subtypes defined by the gene expression data that are more predictive of clinical outcomes (disease recurrence, survival, disease-free survival, etc.) than usual clinical correlates. Development of such a molecular classification system may potentially lead to more tailored therapies for patients as well as better diagnostic procedures.

Hierarchical clustering has been an important tool in the discovery of disease subtypes in microarray data (Eisen et al. 1998). Such procedures typically output a dendrogram that groups samples; an example using the data from the study by Bittner et al. (2000) is provided in Figure 1. Determining the reliability of clustering procedures poses a major problem in the interpretation and analysis of microarray data. It is important to separate the clusters which arise due to random chance from those which represent “true” clusters. A related question is estimating the true number of clusters in a dataset. Several methods have addressed this issue: these include the proposals of Calinski and Harabasz (1974), Hartigan (1975), Krzanowski and Lai (1985), Tibshirani et al. (2001), Ben-Hur et al. (2002) and Dudoit and Fridlyand (2002). In addition, there have been alternative clustering methodologies developed for microarray data (Getz et al., 2000; Ben-Dor et al., 2000). Still more work has been done on assessing the validity of a clustering procedure based on the jackknife (Yeung et al., 2001) and bootstrap methods (Zhang and Zhao, 2001; Kerr and Churchill, 2001).

In most microarray studies, the number of samples profiled is much smaller than the number of genes and ESTs represented on the chip. Due to the number of elements spotted on the microarray, it is reasonable to assume that there is redundant infor-

mation available on them (Xing and Karp, 2001). Consequently, if we cluster samples based on a subset of the spots on the microarray, stable clusters should be replicated on average. This statement heuristically describes our approach to assessing the reliability of clustering analyses of microarray data. We propose calculating cluster stability scores based on subsampling methods. The approach is relatively generic and can be applied to any clustering algorithm. We will focus primarily on hierarchical clustering since that is the technique used most often in the analysis of microarray data. While we focus here on clustering samples, these methods can be utilized for clustering genes as well. These techniques have been examined for supervised learning problems (Ho, 1998); their application to clustering techniques appears to be novel. The issue addressed in this paper is separate from that of estimating the number of clusters in a dataset. However, the two problems are related; in particular, the cluster stability scores depend on the number of clusters. In **System and Methods**, we describe the data used, outline hierarchical clustering and summarize the procedure of Ben-Hur et al. (2002) for estimating the number of clusters. Two approaches are described in this paper. For the first, we assume that the number of clusters is known; cluster stability scores are calculated. In the second situation, the number of clusters is unknown. We address this problem by developing a two-stage procedure in which the number of clusters is estimated at the first stage and sensitivity measures are calculated at the second. These techniques are described in **Algorithms**. We have programmed our procedures in the R language; in **Implementation**, we briefly discuss the software. We use these methods to re-analyze three publicly available datasets in the literature: a B-lymphoma study (Alizadeh et al., 2000), a cutaneous melanoma study (Bittner et al., 2000) and a childhood cancer study (Khan et al., 2001). These analyses are summarized in **Results**. Finally, in **Discussion**, we make some concluding remarks.

Systems and methods

Data and clustering procedures

We will let $\mathbf{x}_1, \dots, \mathbf{x}_n$ denote the p dimensional vectors of gene expression profiles; n is the number of samples profiled. In what follows, we assume that the data have been preprocessed and normalized. Thus, our procedures work with both oligonucleotide and cDNA microarrays.

Since we will be primarily applying our methods to hierarchical clustering procedures, we briefly describe the method here.

Hierarchical clustering

We first constructs a dissimilarity measure for each pair of objects, often a distance measure $d(\mathbf{x}_i, \mathbf{x}_j)$. Alternatively, the dissimilarity measure may be taken to be one minus some measure of association, typically the correlation coefficient ρ .

Hierarchical clustering algorithms begin with n singleton clusters; the closest pair of distinct clusters is found and merged, leaving $(n - 1)$ singleton clusters and one cluster with two distinct objects. We then update the dissimilarity matrix is updated to take into account the merging that has occurred. The two closest distinct clusters are found and merged based on the resulting dissimilarity matrix. We iterate these steps until one cluster remains. There are many ways to update the dissimilarity matrix, the main issue is how to define a distance between two clusters. In average linkage clustering, the distance between two clusters is the average of the pairwise distances between two elements, one from the first cluster and the other from the second. In complete linkage clustering, the distance between two clusters is taken to be the maximum of all possible pairwise distances. At the other extreme is single linkage clustering, where the distance between two clusters is taken to be the minimum of all possible pairwise distances.

Estimating number of clusters

In the **Algorithm** section, we discuss a two-stage procedure for calculating cluster stability scores when the number of clusters is not fixed *a priori*. The method involves estimating the number of clusters at the first stage and then computing the scores at the second stage. We looked at the literature for the various proposals of estimating the number of clusters. Based on our experience with real datasets, the best performance seemed to be given by the method of Ben-Hur et al. (2002). We now briefly describe their procedure. It should be pointed out that our approach is relatively generic and that any method for estimating the number of clusters can be used in the first stage.

In the approach of Ben-Hur et al. (2002), the samples are partitioned into k clusters. We then rerun the clustering algorithm based on the subsampling a fraction of the samples and group the subsamples into k clusters. We then compute a similarity index of the subsamples, the correlation coefficient between the clusters for the resampled data with those for the original data based on the definition given by Fowlkes and Mallow (1983). We repeat this several times to get a histogram of correlation coefficient values. We then vary k and redo the procedure. For values of k where real biological clusters are represented, the histogram of correlation coefficient values will be concentrated around 1. On the other hand, correlation coefficient histograms for larger values of k tend to be spread more uniformly. The estimate for the number of clusters in a dataset is the value of k for which the histograms transition from being concentrated near 1 to being more uniformly distributed.

Algorithm

Cluster stability scores for known number of clusters

In this section, we assume that the number of clusters is known to be some number, say K . Thus, the samples $\{1, 2, \dots, n\}$ are partitioned into K sets A_1, \dots, A_K . To apply

the random subspace, we randomly choose a subset D of the indices $\{1, 2, \dots, p\}$, where the cardinality of D is d . We comment later on the choice of d . We then create a new dataset $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$, where \mathbf{x}_i^* is the d -dimensional subvector of \mathbf{x}_i ($i = 1, \dots, n$). We create a new dissimilarity matrix based on the \mathbf{x}_i^* , $i = 1, \dots, n$ and rerun the hierarchical clustering procedure. The resulting dendrogram is cut into K clusters, A_1^*, \dots, A_K^* . We then check to see if $A_i \subset A_j^*$ for $i, j = 1, \dots, K$. The random subspace selection is repeated B times. For each of the original sets A_1, \dots, A_K , the cluster stability score is defined as the proportion of B samples in which A_i ($i = 1, \dots, K$) appears. If the score is close to 1, then this is evidence that the cluster is stable. On the other hand, if the proportion is small, then the stability of the cluster is less reliable.

These sensitivity measures will depend on the choice of d . Larger values of d tend to yield larger sensitivity measures while the converse holds for small d . Our experience has been to choose d to be within between .75 and .85 times p .

Cluster stability scores for unknown number of clusters

Having developed a method for computing cluster stability scores in the previous section, we now summarize our method when the number of clusters is not known *a priori*. The following two-stage method is adopted. First, we estimate the number of clusters at the first stage using the technique of Ben-Hur et al. (2002) and get an estimate K^* . Next, conditional on K^* , we calculate the cluster stability scores.

Implementation

We are in the process of writing macros in R for implementing the methods we have proposed here. They are obtainable from the first author's website at the following URL:

<http://www.sph.umich.edu/~ghoshd/COMPBIO/CSS/>.

R is a freely downloadable software package (<http://www.r-project.org/>) and can run on either a Windows or UNIX platform.

Results

We now discuss the application of the proposed methodology to three microarray datasets: one from a childhood cancer study (Khan et al., 2001), one from a lymphoma study (Alizadeh et al., 2000) and the final is from a cutaneous melanoma study (Bittner et al., 2000). The results using complete linkage clustering are summarized here; readers can go to the URL mentioned in the previous section to find results based on average linkage clustering.

For each dataset, the algorithm of Ben-Hur et al. (2002) was applied to the hierarchical clustering output. At each iteration of the algorithm, we randomly subsampled 65% of the available samples. In instances for which the true number of clusters was not obvious, both visual inspection of the original dendrogram and examination of the result obtained using the other linkage methods for that dataset were considered.

After estimating the true number of clusters, cluster stability scores were calculated for $d = 85\%$, 75% , 50% and 25% of the total numbers of genes. For each rate, one hundred cluster trials were performed.

In the Khan dataset, gene expression values were measured for 2308 genes on a total of 89 subjects. The dendrogram using complete linkage clustering of these data is presented in Figure 2. For this data, application of the method of Ben-Hur et al. (2002) yielded an estimate of $K = 7$ clusters, the labels of which are listed in Table 1. Their cluster stability scores are presented in Table 2. Based on these results, it does not appear that any of the clusters are highly stable. In fact, cluster 3 appears not to be stable at all. Looking at the tissue labels in Table 1, we see that most of the clusters tend to have distinctly different cancer types grouped together. This is one possible reason for the relatively low cluster stability scores. The other potential reason is that by subsampling spots on the array, we are losing vital information that discriminates the clusters.

In the Alizadeh dataset, data were available on 96 samples for whom gene ex-

pression values on 4026 different genes were measured. The dendrogram based on complete linkage clustering is presented in Figure 3. Application of Ben-Hur et al. (2002) methodology to the complete linkage results suggested the presence of eight true clusters in the data. The groupings for the eight clusters are presented in Table 3. The stability score results are given in Table 4. We see that there is strong evidence that clusters 7 and 8 are highly stable clusters. On the other hand, cluster 2, which is similar to the lymphoma subtype cluster found by Alizadeh et al. (2000), is not stable at all.

Finally, the cluster stability score method is applied to the data from the melanoma study conducted by Bittner et al. (2000). We consider data on 31 samples for whom gene expression measurements on 3613 genes were used. Application of Ben-Hur's method in conjunction with visual inspection yielded an estimate of four true clusters for the complete-linkage clustering solution. The cluster labels and cluster stability scores are presented in Tables 5 and 6, respectively. Again, we see relative low evidence for cluster stability in the four groups with the potential exception of cluster 2.

Discussion

In this paper, we have developed an approach to statistical validation of clustering results based on subsampling methods. One of the advantages of this approach is that it exploits the fact that in microarray experiments, the number of spots on the chip is greater than the number of samples profiled. By subsampling the spots on the chip, we are able to determine which clusters are relatively stable on average. It is important to note that an assumption being made is that there is sufficient correlation on the spots with respect to discriminating between clustered samples. For example, if only one gene on a 10K chip discriminates two cancer subtypes, then the approach described here might give misleading results.

Based on the cluster stability score method, we revisited several datasets from cancer studies to explore the stability of clustered samples. In particular, we found

relatively little statistical evidence for the subtypes found by Alizadeh et al. (2000) and by Bittner et al. (2000). However, our approach is statistical; we did some relatively *ad hoc* biological validation after performing the analyses. In many cancer studies, there are additional clinical covariates (e.g., survival time, PSA recurrence) available. One potential method of more formal biological validation is to combine the clustering methodology with correlation of the subsequent output to these covariates. It is an area we are currently pursuing.

Acknowledgments

This work has been supported by a MUNN Idea Grant and Prostate SPORE Seed Grant from the University of Michigan, as well as a Bioinformatics Pilot Award from the University of Michigan and Pfizer.

References

- Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J. C., Sabet H., Tran T., Yu X., Powell J. I., Yang L., Marti G. E., Moore T., Hudson J., Lu L., Lewis D. B., Tibshirani R., Sherlock G., Chan W. C., Greiner T. C., Weisenburger D. D., Armitage J. O., Warnke R., Staudt L. M. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.
- Ben-Dor A, Shamir R, Yakhini Z. (1999). Clustering gene expression patterns. *Journal of Computational Biology* **6**, 281–297.
- Ben-Hur, A., Elisseeff, A. and Guyon, I. (2002). A stability-based method for discovering structure in clustered data. In *Proceedings of Pacific Symposium on Biocomputing 2002* (Eds. Altman, R. B. et al.). New Jersey: World Scientific Press. pp. 6-17.

- Bittner M., Meltzer P., Chen Y., Jiang Y., Seftor E., Hendrix M., Radmacher M., Simon R., Yakhini Z., Ben-Dor A., Sampas N., Dougherty E., Wang E., Marincola F., Gooden C., Lueders J., Glatfelter A., Pollock P., Carpten J., Gillanders E., Leja D., Dietrich K., Beaudry C., Berens M., Alberts D. and Sondak V. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540.
- Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method to estimate the number of clusters in a dataset. Technical report, Department of Statistics, UC-Berkeley.
- 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.
- Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association* **78**, 553–569.
- Getz G, Levine E, and Domany E. (2000). Coupled two-way clustering analysis of gene expression data. *Proceedings of the National Academy of Sciences* **97**, 12079 – 12084.
- Hartigan, J. (1975). *Clustering Algorithms*. New York: Wiley.
- Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 832–844.
- 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.

- Khan, J., Wei, J. S., Ringnér, Saal, L. H. et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7**, 673 – 679.
- Krzanowski, W. J. and Lai, Y. T. (1985). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics* **44**, 23–34.
- Tibshirani, R., Walter, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Ser B* **63**, 411–423.
- Xing, E. and Karp, R. (2001). CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* **17**, 306S – 315S.
- Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. (2001). Validating clustering for gene expression data. *Bioinformatics* **17**, 309–318.
- Zhang, K. and Zhao, H. (2000). Assessing reliability of gene clusters from gene expression data. *Functional and Integrative Genomics* **1**, 156–173.

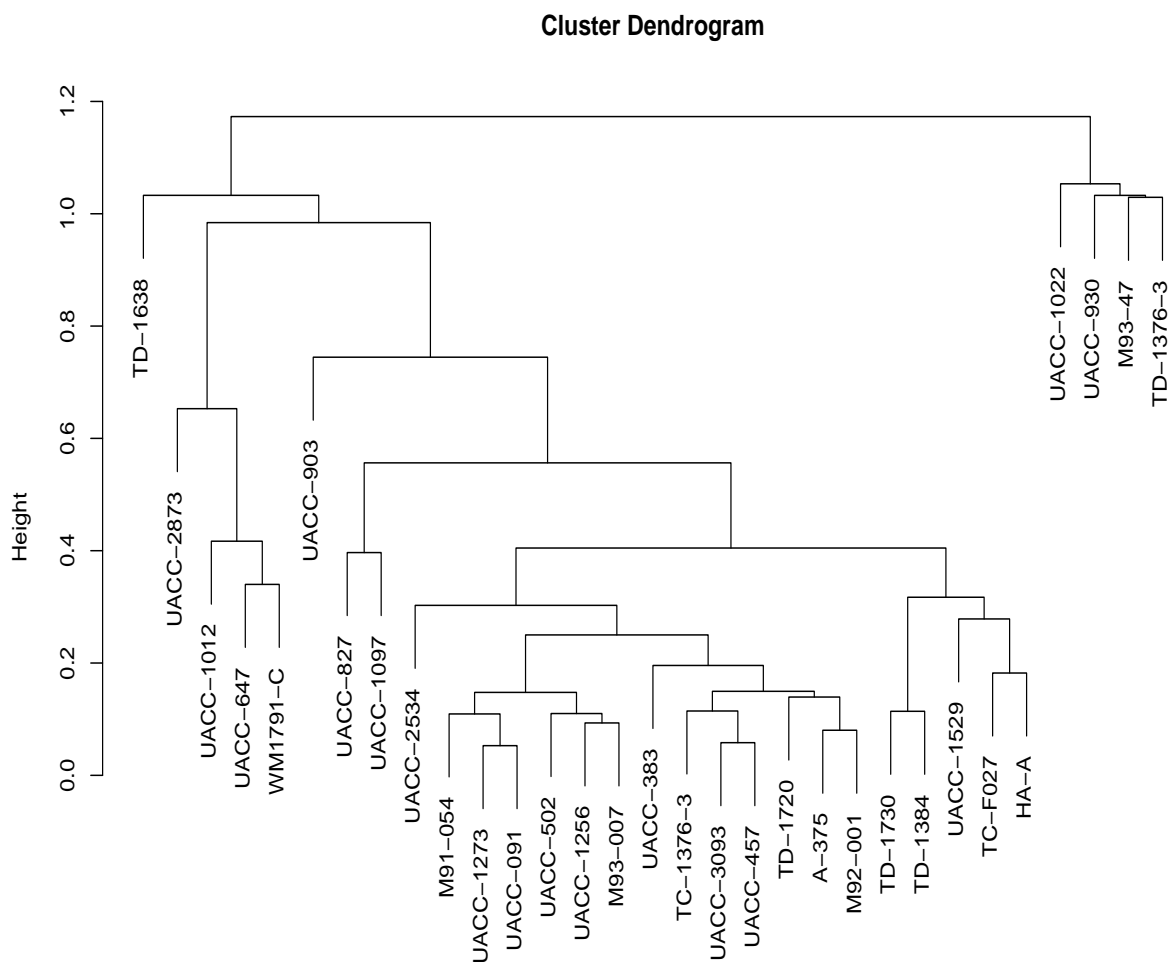


Figure 1: Hierarchical Clustering Dendrogram of gene expression data from Bittner et al. (2000). Average linkage clustering used.

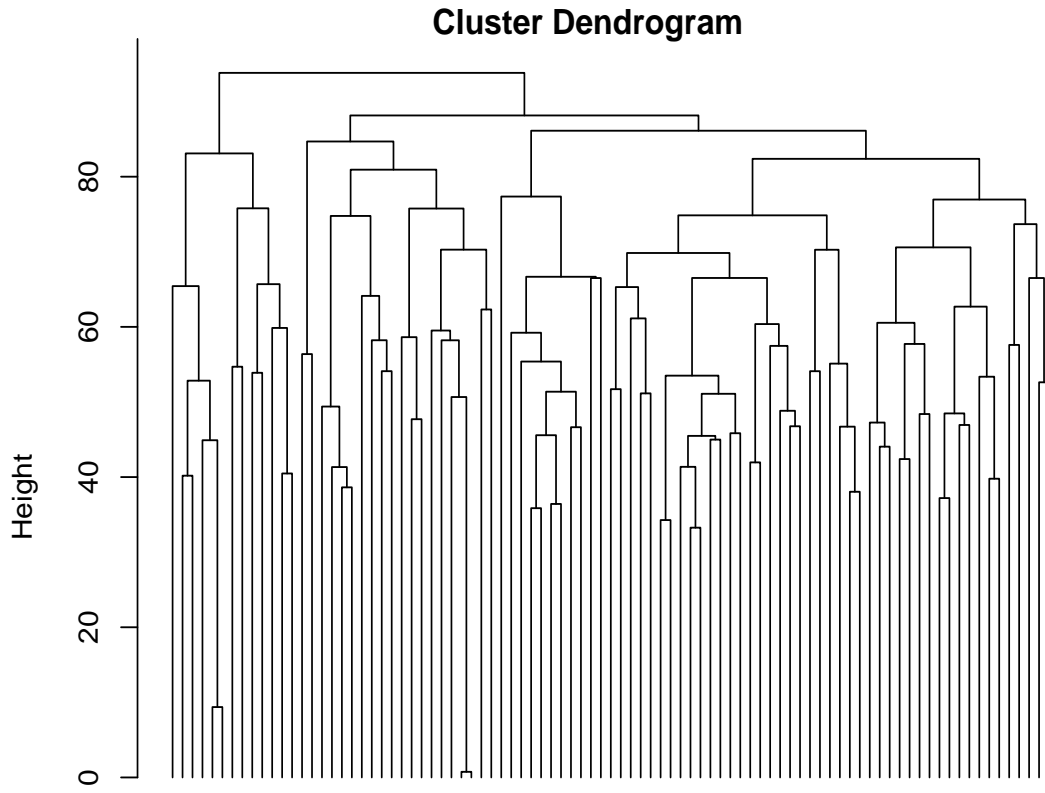


Figure 2: Hierarchical Clustering Dendrogram of gene expression data from Khan et al. (2001). Complete linkage clustering used.

Table 1. Cluster labels for $K = 7$ groups from Khan et al. (2001) data

Cluster	Samples
1	EWS.T1, EWS.T2, EWS.T3, EWS.C3, EWS.C2, EWS.C4, EWS.C1, BL.C1, BL.C2, BL.C3, BL.C4, RMS.C8, RMS.C11, RMS.T1, RMS.T4, RMS.T2, RMS.T3, TEST.5, TEST.24
2	EWS.T4, EWS.T6, EWS.T7, EWS.T9, EWS.T11, EWS.T12, EWS.T14, EWS.T15, EWS.T19, RMS.T5, TEST.6
3	EWS.T13, RMS.C3, RMS.C9, RMS.C5, RMS.T6, RMS.T7, RMS.T8, RMS.T9, RMS.T10, TEST.10, TEST.21, TEST.20, TEST.22, TEST.16, TEST.23, TEST.14, TEST.25, TEST.19
4	EWS.C8, EWS.C6, EWS.C9, EWS.C11, EWS.C10, TEST.3
5	EWS.C7, BL.C5, BL.C6, BL.C7, BL.C8, NB.C1, NB.C2, NB.C3, NB.C6, NB.C12, NB.C7, NB.C4, NB.C5, NB.C10, NB.C11, NB.C9, NB.C8, RMS.C4, RMS.C2, RMS.C6, RMS.C7, RMS.C10, TEST.11, TEST.8, TEST.18, TEST.15,
6	RMS.T11, TEST.1, TEST.2, TEST.4, TEST.7, TEST.12, TEST.17
7	TEST.9, TEST.13

Table 2. Cluster stability scores for childhood cancer data from Khan et al. (2001)

d	Cluster Labels						
	1	2	3	4	5	6	7
85	63	53	4	79	15	67	62
75	61	42	2	71	4	64	60
50	17	5	0	31	1	36	49
25	6	1	0	14	0	21	47

Note: Cluster labels are from Table 1. Complete linkage clustering used.

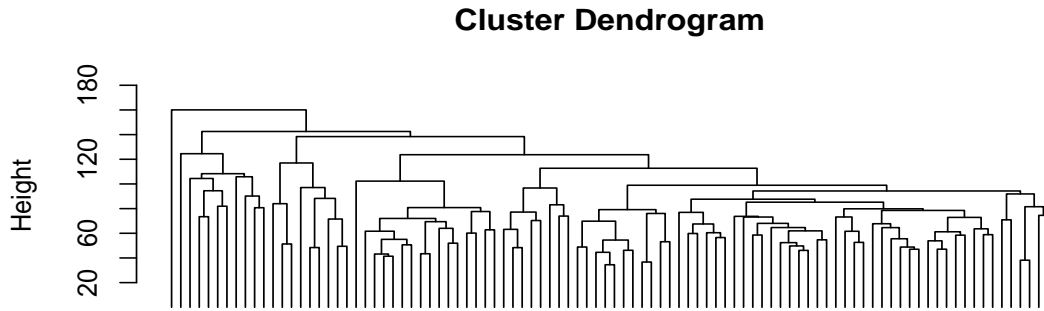


Figure 3: Hierarchical Clustering Dendrogram of gene expression data from Alizadeh et al. (2000). Complete linkage clustering used.

Table 3. Cluster labels for $K = 8$ groups from Alizadeh et al. (2000) data

Cluster	Samples
1	OCI.Ly3, OCI.Ly10, DLCL.0042, Blood.B.cells.anti.IgM.CD40L.24h, Blood.B.cells.anti.IgM.24h, Blood.B.cells.anti.IgM.IL.4.24h, Blood.B.cells.anti.IgM.CD40L.IL.4.24h, Blood.B.cells.anti.IgM.CD40L.6h
2	DLCL.0007, DLCL.0031, DLCL.0036.OCT, DLCL.0030, DLCL.0004, DLCL.0029, DLCL.0008, DLCL.0034, DLCL.0051, DLCL.0011, DLCL.0032, DLCL.0006, DLCL.0049, Tonsil, DLCL.0039, DLCL.0001, DLCL.0018, DLCL.0037, DLCL.0010, DLCL.0015, DLCL.0026, DLCL.0005, DLCL.0023, DLCL.0027, DLCL.0024, DLCL.0013, DLCL.0002, DLCL.0016, DLCL.0020, DLCL.0003, DLCL.0014, DLCL.0048, DLCL.0033, DLCL.0025, DLCL.0040, DLCL.0028, DLCL.0012, DLCL.0021, Blood.B.cells.anti.IgM.CD40L.low.48h, Blood.B.cells.anti.IgM.CD40L.high.48h, DLCL.0009
3	Tonsil.GC.B, Tonsil.GC.Centroblasts, FL.9, FL.9.CD19., FL.12.CD19., FL.10.CD19., FL.10, FL.11, FL.11.CD19., FL.6.CD19., FL.5.CD19.
4	SUDHL6, DLCL.0052, WSU1, DLCL.0041
5	Blood.B.cells.anti.IgM.IL.4.6h, Blood.B.cells.anti.IgM.6h, .anti.IgM.IL.4.6h, Blood.B.cells.anti.IgM.6h
6	Blood.T.cells.Adult.Naive.CD4..Unstimulated, Blood.T.cells.Adult.Naive.CD4..I.P.Stimulated, Cord.Blood.T.cells.Neonatal.Naive.I.P.Stimulated, Blood.T.cells.Neonatal.Naive.CD4..Unstimulated, Thymic.T.cells.Fetal.CD4..Unstimulated, Thymic.T.cells.Fetal.CD4..I.P.Stimulated
7	Jurkat, U937, OCI.Ly12
8	Blood.B.cells.memory.CD27., Blood.B.cells.naive.CD27., Blood.B.cells, Cord.Blood.B.cells, CLL.60, CLL.68, CLL.9, CLL.14, CLL.51, CLL.65, CLL.71.Richter.s, CLL.71, CLL.13, CLL.39, CLL.52

Note: Five other samples (Lymph.Node, DLCL.0017, OCI.Ly1, OCI.Ly13.2, SUDHL5) were grouped as singletons.

Table 4. Cluster stability scores for data from lymphoma study by Alizadeh et al.
(2000)

d	Cluster labels							
	1	2	3	4	5	6	7	8
85	98	19	98	72	99	100	100	100
75	89	10	95	57	92	98	100	98
50	62	8	71	36	75	82	97	88
25	35	3	49	13	53	66	82	72

Note: Cluster labels are from Table 3. Complete linkage clustering used.

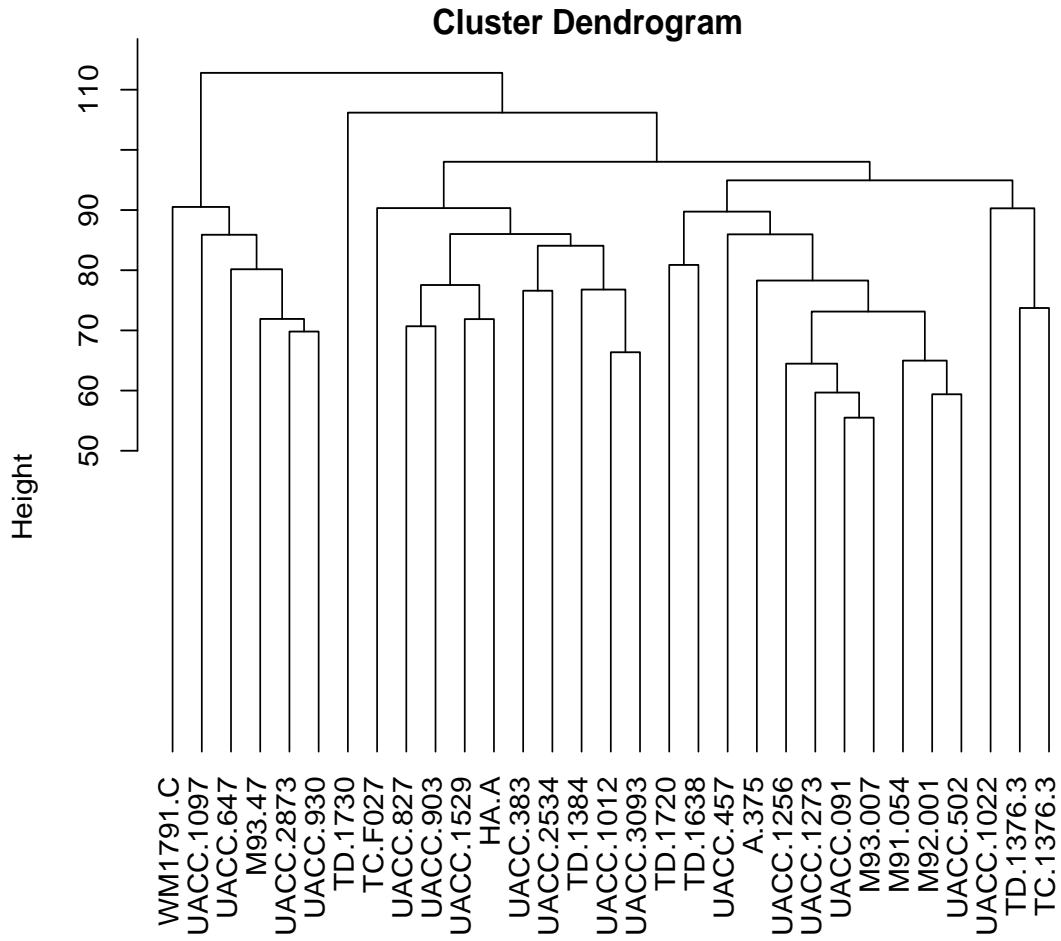


Figure 4: Hierarchical Clustering Dendrogram of gene expression data from Bittner et al. (2000). Complete linkage clustering used.

Table 5. Cluster labels for $K = 4$ groups from Bittner et al. (2000) data

Cluster	Samples
1	TC.F027, UACC.1012, UACC.1529, UACC.827, HA.A, UACC.903, TD.1384, UACC.3093, UACC.383, UACC.2534
2	UACC.2873, UACC.647, WM1791.C, UACC.930, UACC.1097, M93.47
3	TD.1720, TD.1638, A.375, UACC.457, M92.001, UACC.1273, UACC.1256, UACC.502, UACC.091, M91.054, M93.007
4	TD.1376.3, TC.1376.3, UACC.1022

Note: One other samples (TD.1730) was grouped as a singleton.

Table 6. Cluster stability scores for melanoma data from Bittner et al. (2000).

d	Cluster labels			
	1	2	3	4
85	9	98	9	52
75	3	90	4	47
50	3	71	3	34
25	0	48	1	28