

# Assessing local cluster stability in microarray experiments using subsampling methods

Mark Smolkin<sup>1</sup> and Debashis Ghosh<sup>2</sup>

<sup>1</sup> *Division of Biostatistics and Epidemiology, University of Virginia*  
and

<sup>2</sup> *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

**Keywords:** Cancer Studies, Gene Expression, Hierarchical Clustering, Random Subspace, Unsupervised Learning.

## 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 these procedures. While much work has been done on assessing the global question of number of clusters in a dataset, relatively little research exists on assessing stability of individual clusters.

**Results:** 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 algorithm. We propose procedures for calculating cluster stability scores for situations involving both known and unknown numbers of clusters. The methods are illustrated on data from a childhood cancer study (Khan et al., 2001).

**Availability:** Code implementing the proposed techniques can be obtained by contacting the second author.

**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 methods 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 global question pertaining to interpretation of cluster analysis output 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).

Determining the reliability of a given cluster, by contrast, is a local clustering question. Less work has been done in this area (Kerr and Churchill, 2001). However, it is obvious that the global and local questions are related, as the individual clusters will depend on the number of clusters inferred from the dataset.

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 information 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 emphasize the problem of clustering samples in the paper, these methods can be utilized for clustering genes as well. Such techniques have been examined for supervised learning problems (Ho, 1998); their application to clustering or unsupervised learning problems appears to be novel. In addition, we develop a joint procedure for addressing the global and local cluster problems. In **Systems and Methods**, we describe the data used and summarize the procedure of Ben-Hur et al. (2002) for estimating the number of clusters, which is a global clustering question. Two approaches are then described. 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. 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 microarray data from 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*

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. We will be primarily applying our methods to hierarchical clustering procedures, but other methods, such as self-organizing maps, k-means or more recent methods (Getz et al., 2000; Ben-Dor et al., 2000) could be utilized as well.

## *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. In their approach (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. Next, we compute a similarity index of the subsamples, the correlation coefficient between the clusters for the resampled data with those for the original data computed based on the definition given by Fowlkes and Mallows (1983). This is repeated 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.

## **Algorithms**

### *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  $A_1, \dots, A_K$ . We then randomly choose a subset  $D$  of the indices  $\{1, 2, \dots, p\}$ , where  $d$  is the cardinality of  $D$ . A new dataset, consisting of  $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$ , where  $\mathbf{x}_i^*$  is the  $d$ -dimensional subvector of  $\mathbf{x}_i$  ( $i = 1, \dots, n$ ), is then created. We compute 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  $i, j = 1, \dots, K$ . This resampling 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.

A parameter in the procedure is  $d$ . The stability scores 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  $p$ .

The sensitivity measure computed here is an estimate of a probabilistic quantity that is averaged over  $B$  models, where each model is based on a random subset of the data. This provides an analogue of stacking or combining models (Wolpert, 1992) for unsupervised learning. It might be also possible to calculate sensitivity measures that average both over  $d$  as well as over subsets of  $(1, \dots, p)$ , but we will not pursue that here.

### *Cluster stability scores for unknown number of clusters*

In the previous section, we developed the calculation of cluster stability scores in the case where the number of clusters is known. If, on the other hand, the number of clusters is not known, then this has to be estimated somehow. We propose the following two-stage method:

1. We estimate the number of clusters at the first stage using the technique of Ben-Hur et al. (2002) and get an estimate  $K^*$ .
2. Conditional on  $K^*$ , we calculate the cluster stability scores.

Observe that any method for choosing number of clusters, such as those listed in the introduction, could be used in step 1 of the procedure.

### **Implementation**

We are in the process of writing macros in R for implementing the methods we have proposed here. When ready, they will be obtainable from the second author's website at the following URL:

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

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

### **Results**

We applied 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 last from a cutaneous melanoma study (Bittner et al., 2000). Because of space limitations, the results from the last two can be found at the second author's website, the URL for which was given in the previous section. For implementation of the Ben-Hur et al. (2002) 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 subsamples were generated.

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. The cluster stability scores are presented in Table 2. Based on these results, the most stable cluster was cluster 4, which consisted of Ewing’s sarcomas; based on the original paper, the test sample is an osteosarcoma. The next set of stable clusters are clusters 1, 6, and 7. As reported in the original paper, cluster 7 consists of two normal muscle tissue samples.

As was mentioned before, the cluster stability scores depend on  $d$ . As  $d$  decreases, the scores decrease as well. Based on the results of Table 2, the one cluster that remains stable for varying values of  $d$  is cluster 7. However, note that the relative rankings of the clusters appears to be unchanged. We suggest using the cluster stability scores as a relative measure rather than as an absolute one.

## Discussion

In this paper, we have developed a simple 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 reanalyzed several datasets from cancer studies to explore the stability of clustered samples. In particular, we found relatively little statistical evidence to support the claims of subtypes found by Alizadeh et al. (2000) and by Bittner et al. (2000). However, our approach is statistical and should by no means serve as a substitute for experimental validation.

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.

## 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. *Genome Biology* **3**, RESEARCH0036.1 – 0036.21.
- 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.
- Wolpert, D. (1992) Stacked generalization. *Neural Networks* **5**, 241–259.
- Xing, E. and Karp, R. (2001). CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* **17**, 306S – 315S.

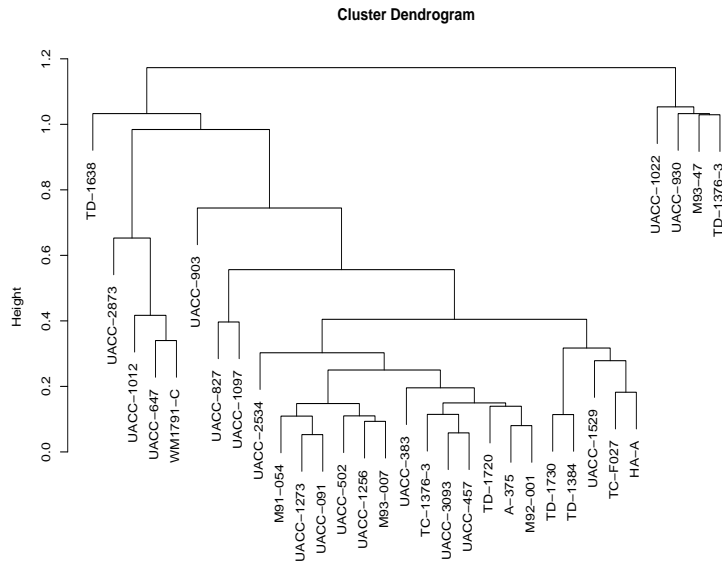


Figure 1: Hierarchical Clustering Dendrogram of gene expression data from Bittner et al. (2000). Average linkage clustering used.

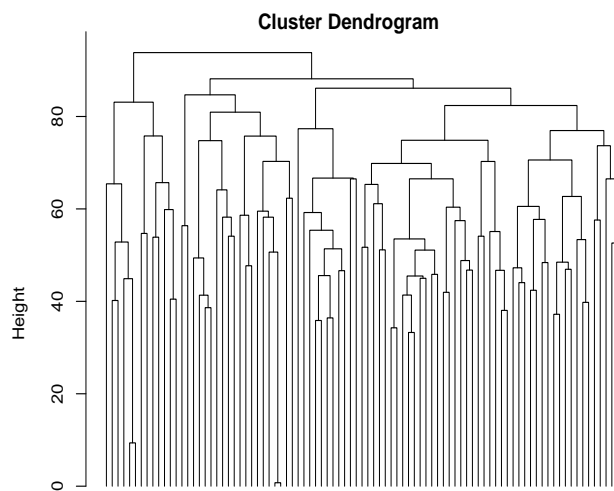


Figure 2: Hierarchical Clustering Dendrogram of gene expression data from Khan et al. (2001). Complete linkage clustering used.

**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)

$d/2308$	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.