Statistica Sinica **12**(2002), 219-240

STATISTICAL ISSUES IN THE CLUSTERING OF GENE EXPRESSION DATA

Darlene R. Goldstein, Debashis Ghosh and Erin M. Conlon

University of California, University of Michigan and Harvard University

Abstract: This paper illustrates some of the problems which can occur in any data set when clustering samples of gene expression profiles. These include a possible high degree of dependence of results on choice of clustering algorithm, further dependence of results on the choices of genes and samples to be included in the clustering (for example, whether or not to include control samples), and difficulty in assessing the validity of the grouping. We also demonstrate the use of Cox regression as a tool to identify genes influencing survival.

Key words and phrases: Cluster analysis, Cox regression, microarray experiment, survival analysis, unsupervised learning.

1. Introduction

With the advent of complementary DNA (cDNA) microarray technology (Schena, Shalon, Davis and Brown (1995); Schena (1999)), it has become possible to simultaneously measure the messenger RNA (mRNA) levels for thousands of genes and expressed sequence tags (ESTs) in parallel. This capability has led to consideration of genomic data for organisms on a large-scale basis. The use of microarrays has led to an explosion of molecular profiling studies. In these studies, samples of cells in different experimental conditions are analyzed using cDNA microarrays. The measured mRNA levels on the microarrays are subsequently compared across samples in order to determine which genes are differentially regulated under certain conditions. This helps us begin to understand potential mechanisms of transcription regulation. Settings where molecular profiling studies have occurred include cell cycle determination in yeast (Spellman et al. (1998)) and studies of acute leukemias (Golub et al. (1999)), lymphomas (Alizadeh et al. (2000)), and breast cancers (Perou et al. (1999)).

A seminal paper in the analysis of microarray data is that of Eisen, Spellman, Brown and Botstein (1998), in which the authors propose hierarchical clustering of genes as a means to identify patterns in the high-dimensional data generated by microarrays. Clustering of samples may also be performed; even two-way clustering of genes into functional groups, and of samples into classes, based on gene expression has been done (Alon et al. (1999)). It is now commonplace for

researchers to perform a hierarchical clustering of microarray data to identify patterns in the clustering. In many instances, cluster analysis is the primary technique of data analysis, regardless of the specific questions of interest.

Fundamental to any analysis of data is the scientific question of interest. It is the question to be addressed that ought to guide the choice of analysis tools and techniques. Unfortunately in microarray studies the question may not be carefully posed, may be implied but unreported, or may even seem to be completely lacking. A paper may report, for example, that the aim is to discover subclasses of tumor types. Thus, cluster analysis would seem to be appropriate. However, once the samples are clustered, the investigators reveal that survival data are available; the next move is then an attempt to relate survival to the newly found clusters so that genes influencing survival may be uncovered. If the goal is to find genes influencing survival though, then methods which explicitly handle response variables are more appropriate. In this case, the clustering of samples would seem to be unnecessary.

This paper considers clustering of samples based on gene expression profiles. Our major aim is to highlight some of the issues that arise with the use of hierarchical clustering techniques in the analysis of cDNA microarray data. To illustrate these issues, we include analysis of data from a study of cutaneous melanoma gene expression (Bittner et al. (2000)). We do not present a careful re-analysis of this data set, but rather use it to highlight issues that arise in cluster analysis of tumor profiles. Although we restrict our attention to the spotted array technology, similar issues arise for oligonucleotide arrays (Lockhart et al. (1996); Lipshutz, Fodor, Gingeras and Lockhart (1999)) as well.

The overview of the paper is as follows. In Section 2, we give a brief survey of clustering methodologies. In Section 3, we describe the melanoma cDNA microarray data analyzed by Bittner et al. (2000). In Section 4, statistical issues involved with the application of clustering techniques are given. Because the Bittner et al. (2000) data set has been made publicly available, we are able to demonstrate these issues with it, as well as test alternative analyses. Bear in mind, however, that these same issues may well arise in clusterings of other data sets, yet remain hidden due to the secrecy of the primary data. In Section 5, we introduce a method to identify genes influencing survival; we conclude with a brief discussion in Section 6.

2. Clustering Methods

Here is a brief introduction to several clustering techniques; more comprehensive accounts can be found in Jain and Dubes (1988), Kaufman and Rousseeuw (1990), Everitt (1993) and Gordon (1999). We do not discuss model-based clustering methods; the interested reader is referred to Banfield and Raftery (1993).

We assume in this section that there are n samples or objects to be clustered; each object is associated with a vector of attributes $\mathbf{x}_i = (x_{ij})$, $i = 1, \ldots, n$; $j = 1, \ldots, p$. To avoid issues of scaling, it is further assumed that for $i = 1, \ldots, n$, $\mathbf{x}_i$ is standardized to have mean zero and variance one.

2.1. Hierarchical clustering

To implement the standard method for the analysis of gene expression data from microarray experiments, one first constructs a dissimilarity measure for each pair of objects, often a distance measure $d(\mathbf{x}_i, \mathbf{x}_j)$. Table 1 gives some examples. Alternatively, the dissimilarity measure may be taken to be one minus some measure of association, typically the correlation coefficient ρ.

Table 1. Distance measures used for hierarchical cluster analysis.

Name	$d(\mathbf{x}_i, \mathbf{x}_j)$
Euclidean	$\{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\}^{1/2}$
Manhattan	$\sum_{k=1}^{p} \lvert x_{ik} - x_{jk}\rvert$
Canberra	$\sum_{k=1}^{p} \frac{\lvert x_{ik} - x_{jk}\rvert}{x_{ik} + x_{jk}}$
Maximum	$\max_{1 \le k \le p} \lvert x_{ik} - x_{jk}\rvert$

Hierarchical clustering methods fall into two classes: agglomerative nesting methods and divisive analysis methods (Kaufman and Rousseeuw (1990)). Agglomerative nesting algorithms proceed in the same general manner: 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; the dissimilarity matrix is updated to take into account the merging that has occurred; based on the new dissimilarity matrix, the two closest distinct clusters are found and merged; iterate until one cluster consisting of all n objects 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. Other types are centroid clustering, median clustering and Ward's hierarchical clustering (Ward (1963)). These six types of agglomerative hierarchical clustering algorithms can be shown to be special cases of a general framework considered by Lance and Williams (1967).

The opposite to agglomerative nesting is a divisive analysis approach. Heuristically, the algorithm begins with one cluster of n objects. The object in the cluster that has the greatest dissimilarity to the other elements (the seed) is then separated to form a so-called splinter group and the remaining elements in the original cluster are examined to see whether or not additional elements should be added to the splinter group. Two clusters result. The diameter of each cluster (the largest distance between observations in the same cluster) is then computed to see which one is greater. The steps above are repeated with the cluster that has the greater diameter. Iterate until there are n singleton clusters. The distance for separate clusters can be defined based on average linkage or one of the other methods described above.

A major drawback of hierarchical clustering methods is that group assignment of objects cannot change once an object has been placed in a cluster. These methods cannot undo what has been done in previous steps. In contrast, partition methods can reconsider cluster assignments at every stage.

2.2. Partition clustering

The clustering methods just described are termed hierarchical, and the implied hierarchy can usually be visually represented as a dendrogram. However, there is an important class of clustering methods that are non-hierarchical in nature. These are known as partitioning methods.

Let $C_K = \{c_1, \ldots, c_K\}$ denote a set of K cluster centers and $c(\mathbf{x}) \in C_K$ denote the center closest to $\mathbf{x}$ with respect to some distance measure. The clustering problem amounts to minimizing $\sum_{i=1}^{n} d\{\mathbf{x}_i, c(\mathbf{x}_i)\}$ over C_K. In words, the goal is to find a set of centers such that the mean distance of an object to the closest center is minimized. Algorithms such as K-means clustering (MacQueen (1967)) and partitioning around medoids, or PAM (Kaufman and Rousseeuw (1990)), fall into this class. What distinguishes K-means and PAM is that cluster centers in K-means are averages of objects, while in PAM the centers are actual objects.

It is important to note that these procedures require that K be specified in advance. The problem of choosing the number of clusters K is not a trivial one, and in fact there have been many proposed solutions. Some of these are touched on in Section 4.

3. Cutaneous Melanoma Data Set

In this section, we briefly describe the cDNA microarray experiment and analyses conducted by Bittner et al. (2000). The authors of this study attempted to determine whether or not molecular profiles generated by cDNA microarrays could be used to identify distinct subtypes of cutaneous melanoma, a malignant neoplasm of the skin.

3.1. Study sample

The data consisted of 38 samples from both tissue biopsies and tumor cell lines, including 31 cutaneous melanomas and 7 controls. Two pairs of samples were taken from the same patients. The samples were from both male and female patients aged 29 to 75, with 3 patients of unknown age. The samples were taken from biopsy sites categorized as skin/external, internal and lymph nodes. The controls were of several cell line types including fibroblast, ovarian adenocarcinoma and cell culture variants.

3.2. Hybridization and data preparation

The mRNA was extracted and Cy5-labeled cDNA was created for the 31 cutaneous melanoma and 7 control samples. A single reference probe, labeled Cy3, was used for all 38 samples. The Cy5 and Cy3-labeled cDNA was mixed for each sample and hybridized to a separate melanoma microarray.

The melanoma array contained 8,150 human cDNAs, of which 6,912 were sequence verified under a Cooperative Research and Development Agreement with Research Genetics. The 8,150 cDNAs represented 6,971 unique genes, based on the Unigene build of March 9, 2000. The hybridized array was scanned using both red and green lasers, and the resulting image was analyzed. Of the 8,150 cDNAs, 3,613 were identified as well measured, by having average fluorescence intensity levels above background across all experiments greater than 2,000 (measured on a 16-bit scale) for the less intense signal (Cy3 or Cy5), and by having average spot size greater than 30 pixels for all experiments.

Expression ratios of Cy5/Cy3, or red/green (R/G), were calculated for the 3,613 well-measured genes. These expression ratios are publicly available from the web site http://www.nhgri.nih.gov/DIR/Microarray/Melanoma_Supplement/index.html. The ratios of R/G can become too large or too small when the signal from one source is large and the other is undetectable. The study had many ratios above 10,000 and many below 0.02. Ratios greater than 50 and less than 0.02 were truncated to 50 and 0.02, respectively. The resulting ratios were transformed to a logarithm scale (base 2). The log-ratios were normalized by subtracting the median log-ratio within an experiment from all log-ratios for that experiment, so that the median log-ratio within an experiment is zero. No normalization was performed across experiments, since a single reference probe was used for all experiments.

3.3. Hierarchical clustering

The remaining analyses were performed by Bittner et al. (2000) on the 31 cutaneous melanoma samples, excluding the 7 control samples. Average linkage hierarchical clustering was carried out on the 31 cutaneous melanoma samples.

Thus a matrix was made up of pairwise Pearson correlation coefficients of log-ratios for all experiments and dissimilarity was taken to be 1 - correlation. We replicated their result and give the resulting dendrogram in Figure 1.

If the dendrogram is cut at a height (1 – correlation) of 0.54, the set of 19 samples at the center (numbers 13 – 31) form what Bittner et al. (2000) refer to as the "clustered" melanoma samples (cluster 1); the other 12 samples (1 – 12) fall into several smaller clusters and are referred to as "unclustered" (cluster 2). There seems to be no objective criterion by which a cutoff of 0.54 was chosen; rather, the cutoff seems to result from viewing the dendrogram.

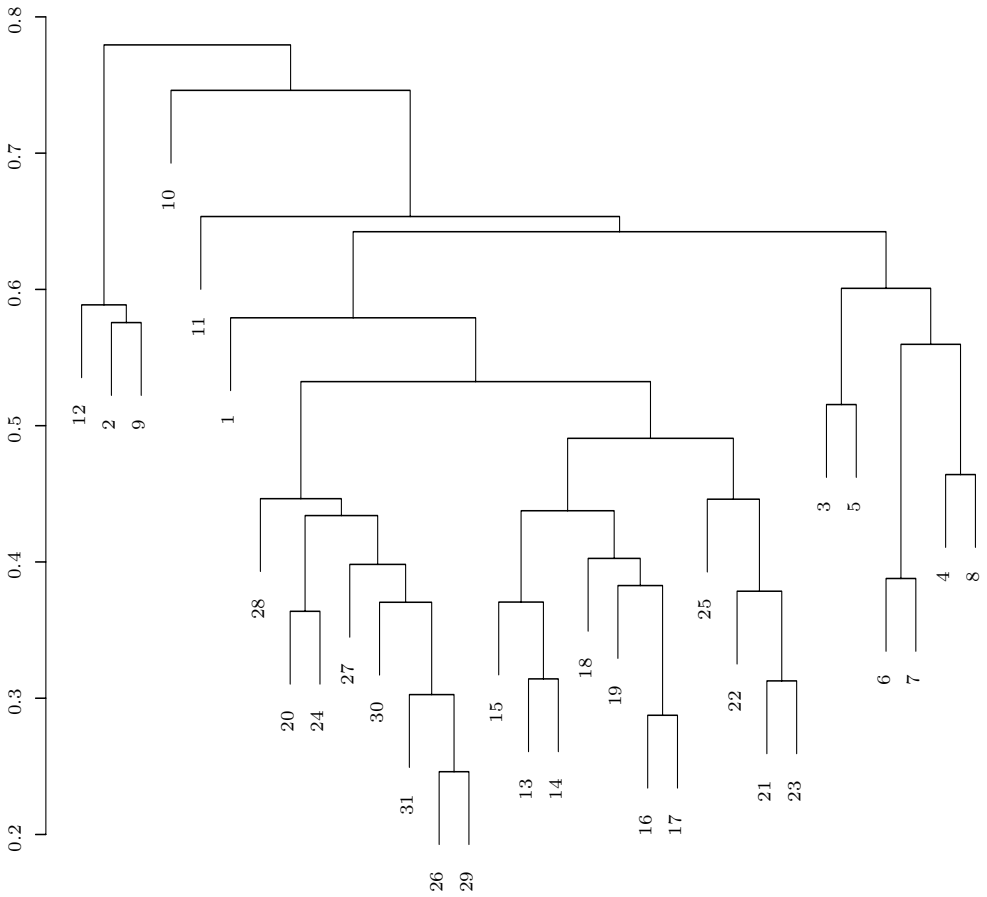

Figure 1. Hierarchical Clustering Dendrogram of Melanoma samples. Average linkage with correlation-based distance.

3.4. Multidimensional scaling and cluster affinity search technique

Additional analyses were performed by Bittner et al. (2000) in an attempt to confirm this clustering. These included multidimensional scaling (MDS) and

a non-hierarchical clustering algorithm, CAST (Ben-Dor, Shamir and Yakhini (1999)). CAST uses a graph theory algorithm for clustering that requires neither a similarity function nor a pre-specified number of clusters. It does use a Bayesian method for choosing the number of clusters and so requires a prior distribution for this quantity. Both the MDS and CAST procedures identified the same major cluster of 19 as that found by average linkage hierarchical clustering.

3.5. Discriminating genes

After identifying and verifying the two clusters, the authors wished to determine whether the samples in the two clusters had different metastatic ability. The study introduced an ad hoc weighting scheme (described in the web supplement to their paper) to score genes based on the gene's ability to distinguish between the two clusters. The weights are ratios of the between cluster Euclidean distance to the weighted average (across pairs) of within cluster distance: $w = d_B/(k_1 d_{w_1} + k_2 d_{w_2} + \epsilon)$, where d_B is the distance between cluster centers, d_{w_i} is the average (Euclidean) distance among all sample pairs (total numbers n_1 and n_2 for clusters 1 and 2, respectively), and $k_i = n_i/(n_1 + n_2)$. A small constant ϵ is added to prevent a zero denominator. Conditional on the cluster assignment, the weight (without the ϵ) is like a t-statistic with an incorrect denominator; this denominator is a sum of square roots rather than the square root of a sum of squares.

Genes were weighted assuming two clusters of 19 and 12 samples. The weight function was designed so that a higher weight corresponded to a greater ability of a gene to distinguish between the two clusters; higher weight corresponded to more compact clusters and greater distance between clusters.

The weight function was calculated for each of the 3,613 genes. There were 182 genes in the weighted gene list that could distinguish between the two clusters, with weights ranging from 6.20 to 0.73. The weighted gene list identified genes that were most differentially expressed in the two clusters. This gene list was used by Bittner et al. (2000) to compare gene expression in samples with known invasive properties to the cutaneous melanoma samples. Uveal melanoma samples were chosen for comparison purposes. It was reported by Bittner et al. (2000) that highly expressed genes in the invasive uveal samples showed the opposite expression pattern of the same genes in the major cluster of 19 cutaneous melanoma samples for genes in the weighted gene list. Downregulated genes for the major cluster of 19 include those genes involved in tumor spreading and migration. This finding suggests reduced motility and reduced invasive ability of the samples in the cluster of 19.

Five specific genes with reduced expression in the cluster of 19 were identified: integrin $\beta 1$, integrin $\beta 3$, integrin $\alpha 1$, syndecan 4, and vinculin. The study further

examined three properties of the samples: ability to migrate into scratch wounds, contract collagen gels, and form tubular networks. Samples within the major cluster of 19 were found to have reduced motility, reduced invasive ability and reduced vasculogenic mimicry, which indicates that these 19 samples may have a less invasive type of cutaneous melanoma, and therefore possibly increased survival.

3.6. Clinical and tumor cell characteristics

Bittner et al. (2000) carried out a number of statistical tests in an effort to determine association between cluster and the following clinical and tumor cell characteristics: sex, age, p16 mutation status, biopsy site, pigment, Breslow thickness, Clark level and specimen type. None of these variables showed an association with cluster, although age had a marginal significance level. The study concluded that the characteristics examined were not sufficient to identify classes of cutaneous melanoma, thus "confirming" the need for gene expression profiling.

3.7. Survival analysis

To further confirm the different metastatic properties of the two groups of cutaneous melanomas, a Kaplan-Meier survival analysis was performed on the 15 (out of 31) samples for which survival data were available. Three deaths occurred among the ten patients classified as being in the "cluster", and there were four deaths among the five patients in the remaining ("unclustered") group. Their survival plots indicate better survival for the cluster of 19, suggesting a less invasive type of cutaneous melanoma for this group (Figure 9). The result is not statistically significant ($p - value = 0.135$), which may be due to a small sample size ($n = 15$) and event rate (7 observed deaths).

3.8. Study conclusions

The study by Bittner et al. (2000) used hierarchical clustering to identify two groups of cutaneous melanomas with different gene expression profiles, the "clustered" group of 19 melanomas (samples 13 – 31) and the "unclustered" group of 12 melanomas (samples 1 – 12). Further investigation identified differing invasive qualities of the cutaneous melanomas in the two groups. The study concluded that melanoma is a useful model in identifying genes important in tumor metastasis.

4. Statistical Issues

We now describe some statistical issues involved in the analysis of the microarray data from Bittner et al. (2000), again noting that these issues are not unique to this particular data set.

4.1. Clustering algorithm choice

We first sought to explore the effects of different clustering algorithms. Several points are noted based on this set of analyses. First, the results of the various clustering algorithms yield different sets of clusters that group together. For example, while average linkage cluster analysis yielded a cluster of 19 melanoma samples, a cluster analysis using complete linkage yields a cluster of 22 melanomas, as opposed to 19 (samples 1, 4, and 8 are now included in the cluster). The dendrogram is shown in Figure 2.

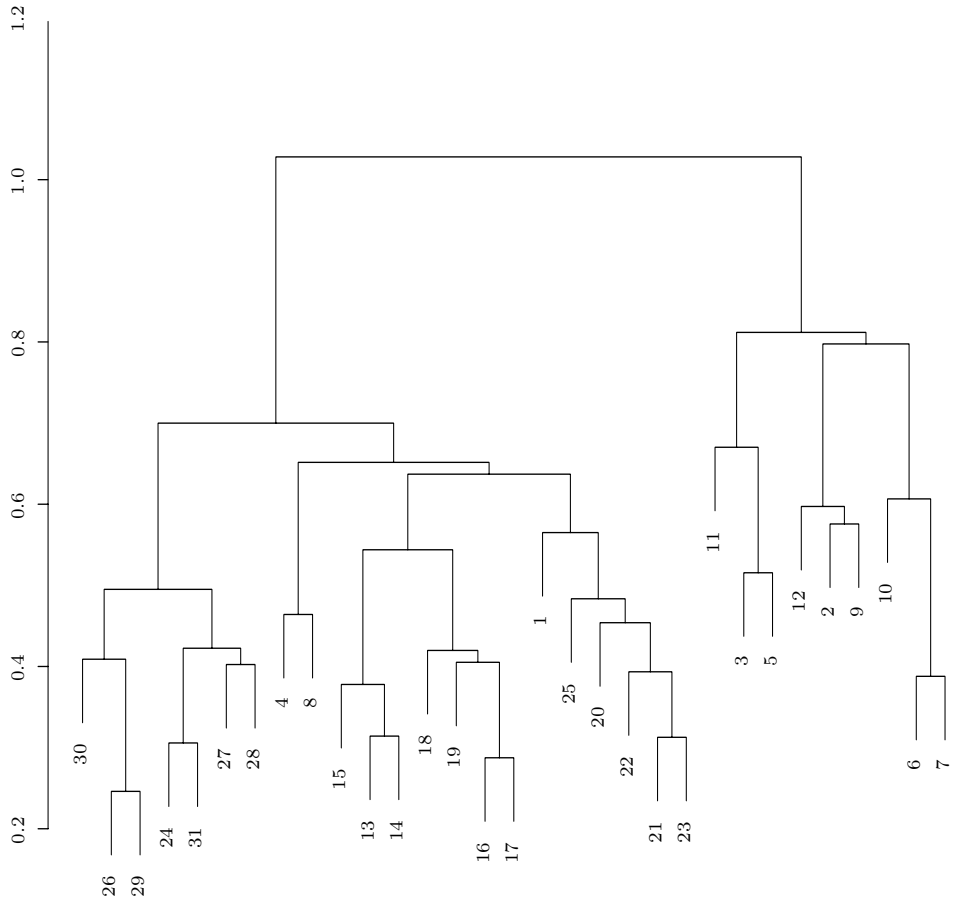

Figure 2. Hierarchical Clustering Dendrogram of Melanoma samples. Complete linkage with correlation-based distance.

A second point to note is that even in situations when a cluster of 19 melanomas was found, the cutoff value for the dendrogram changed from the 0.54 original choice of Bittner et al. (2000). As an example, let us consider fit-

ting the data using the single linkage method for cluster analysis. The resulting dendrogram can be found in Figure 3. The cluster of 19 melanomas can be found if a 1 - correlation cutoff value of about 0.42 is used. If the original cutoff of 0.54 is applied, a cluster of 27 melanomas is obtained. There is not a standard criterion or algorithm for choosing such a cutoff point for a dendrogram. Rather, this choice is often made by visual inspection. Thus, there is some subtlety in the use of dendrograms in order to determine true clusters.

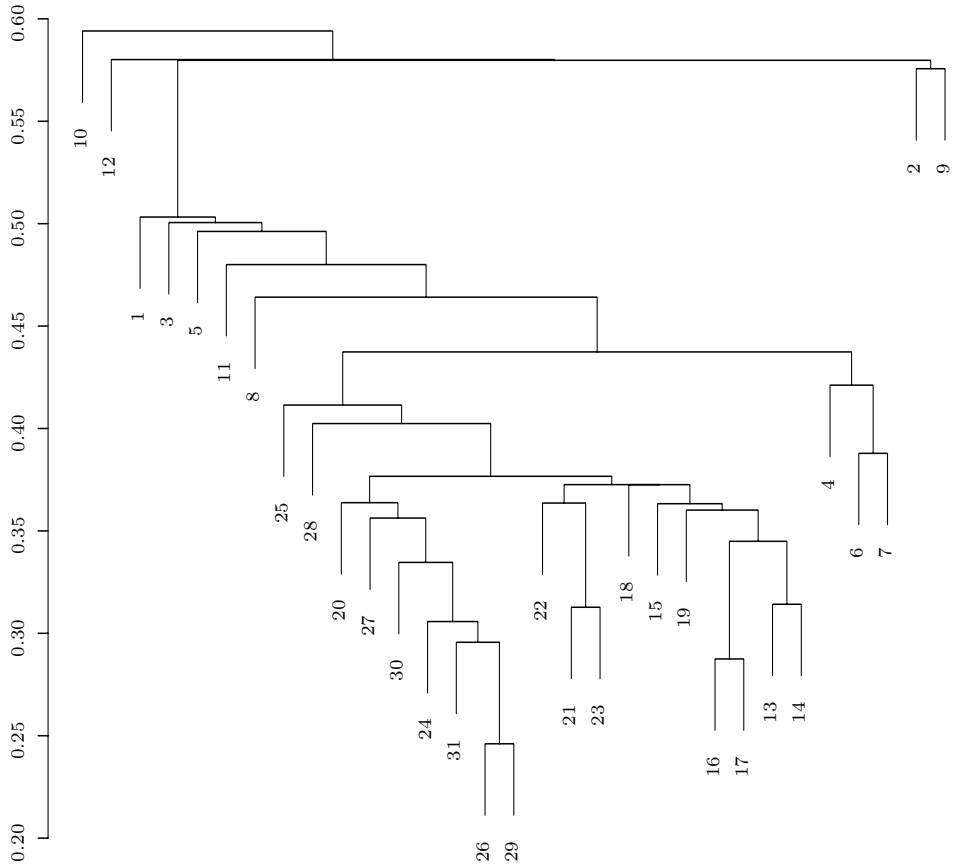

Figure 3. Hierarchical Clustering Dendrogram of Melanoma samples. Single linkage with correlation-based distance.

When divisive hierarchical clustering is used (again, with dissimilarity 1 - correlation), the picture changes more noticeably (Figure 4). The "cluster" of 19 is no longer distinguished, it has been joined by samples 1, 4, 7, 8, and 11.

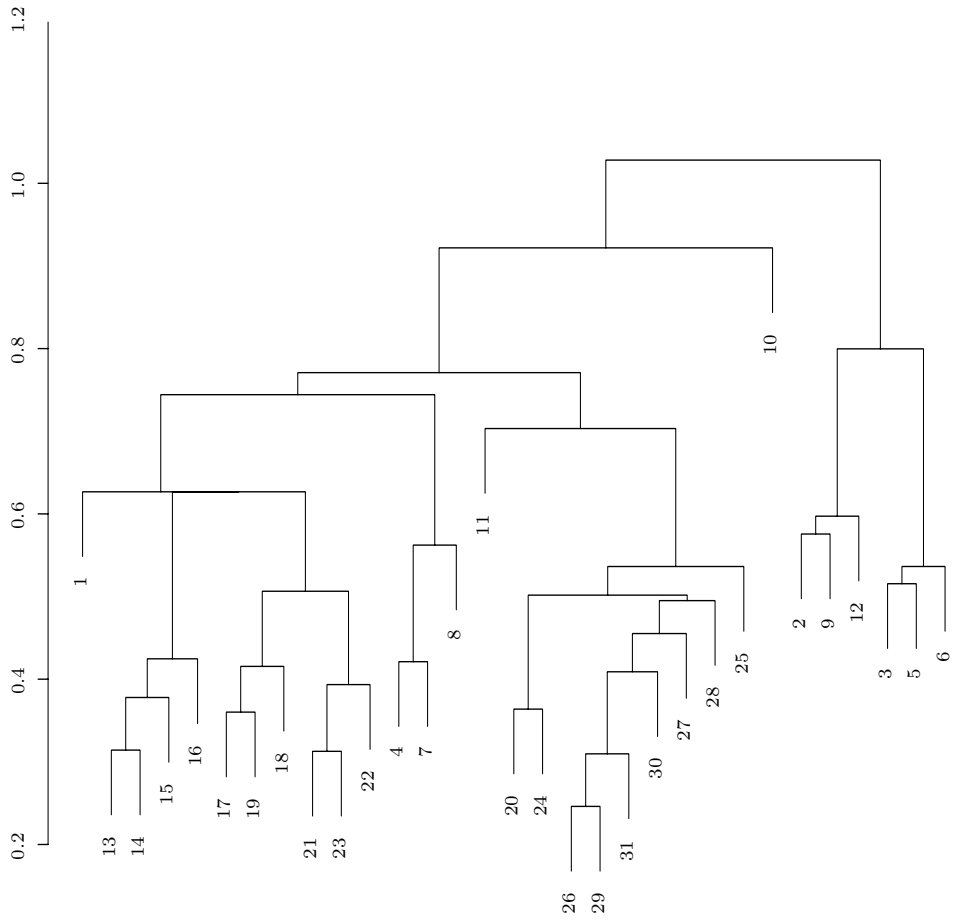

Figure 4. Hierarchical Clustering Dendrogram of Melanoma samples. Divisive Clustering with correlation-based distance.

Partitioning methods yield still further differences in the cluster groupings. We did not attempt to estimate K, but rather took K fixed at 2. The aim here was not to use the data to estimate the true number of clusters, but rather to look at the effect of clustering algorithm on the assignment of samples into 2 groups. The group assignment of Bittner et al. (2000), based on average linkage (agglomerative) hierarchical clustering, can be considered as:

Cluster 1: 13 – 31 (19 members)

Cluster 2: 1 – 12 (12 members)

We applied both K-means and PAM. Table 2 gives the number of samples assigned to each possible combination of clusters across methods. For example, there were 10 samples assigned to cluster 1 in each method, none assigned to cluster 1 by both Bittner et al. (2000) and K-means but to cluster 2 by PAM,

etc. It can be seen by adding the appropriate row totals that the K-means and PAM procedures have a higher level of agreement than either method with the Bittner et al. (2000) assignment.

Table 2. Cross-tabulation of cluster assignments.

Bittner	K-means	PAM	Number of Samples
1	1	1	10
1	1	2	0
1	2	1	1
1	2	2	8
2	1	1	1
2	1	2	0
2	2	1	6
2	2	2	5

These results show that group assignment depends crucially on the choice of clustering algorithm, and also indicates that the assignment of the 19 samples to a single cluster is not as reliable and reproducible as suggested by Bittner et al. (2000).

4.2. Choice of genes and samples for clustering

Perhaps an even more fundamental issue than which particular clustering algorithm to use is which samples will be clustered in the first place, and on which variables (genes). The arrays contained 8,150 cDNA spots; however, only 3,613 of these, considered to be well-measured (sufficiently high compared to background across all samples), were used in the Bittner et al. (2000) analysis. Yet it seems possible that those genes with low measured expression levels (close to the background level) in some arrays are just not being transcribed in that tissue. Finding genes which are differentially expressed across arrays is a major aim of microarray analysis. Genes with very low expression levels in some samples but not in others could be expected to be the basis of an unknown subclass of tumors; their removal from the data set hides a truth (contributes to false negative results), and could conceivably also encourage spurious (false positive) results. We were unable to examine the effect of including genes with very low expression levels on clustering results because the publicly accessible data set contains only the ratios of intensities, and not the separate (red and green) intensities.

The data set does include microarray data on 7 control samples in addition to the 31 melanoma samples. However, these are not used in their clustering. We were therefore interested to see how inclusion of these samples would affect the clustering process.

One potential problem with interpretation of these results, and perhaps a reason the controls were not used by Bittner et al. (2000), is that the control samples were taken from a variety of tissues and biopsy sites. Thus, the inclusion of the mixture of tissue types introduces additional noise and it could be this variation that the clustering picks up. However, the 31 melanoma samples that Bittner et al. (2000) cluster also differ with respect to tissue type and biopsy site, so that even their original clustering may at least in part be affected by this variation.

With agglomerative, average linkage (1-correlation) hierarchical clustering, the 19 "clustered" melanoma samples (13 – 31) again appear together as a single cluster for a cutoff around 0.54 (Figure 5). However, the control samples (32 – 38) do not cluster together, but rather appear interspersed with the 12 "unclustered" melanoma samples (1 – 12). When single linkage is used, the details change but the overall structure is similar to that obtained with average linkage, although the cutoff is changed.

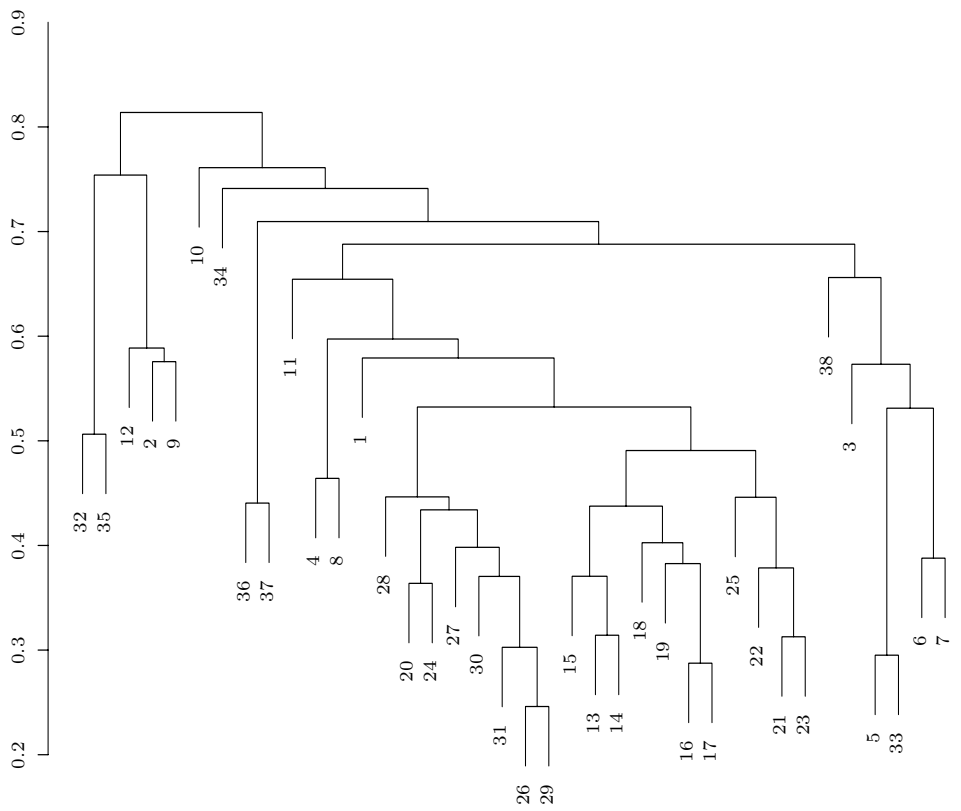

Figure 5. Hierarchical Clustering Dendrogram of Melanoma and Control samples. Average linkage with correlation-based distance.

With complete linkage, the same three "unclustered" samples (numbers 1, 4, and 8) join the major cluster, while two other clusters are apparent (not shown). These two clusters, however, do not correspond to controls and "unclustered" samples; rather, each cluster contains both types of samples.

When all (melanoma and control) samples are clustered with a divisive hierarchical algorithm, the major cluster appears more intact than it did when divisive clustering was used on only the melanoma samples (Figure 6). In this case, 18 of the 19 cluster together, with only sample number 18 clustering away from the group. However, again, the remaining ("unclustered") melanoma samples do not cluster separately from the control samples.

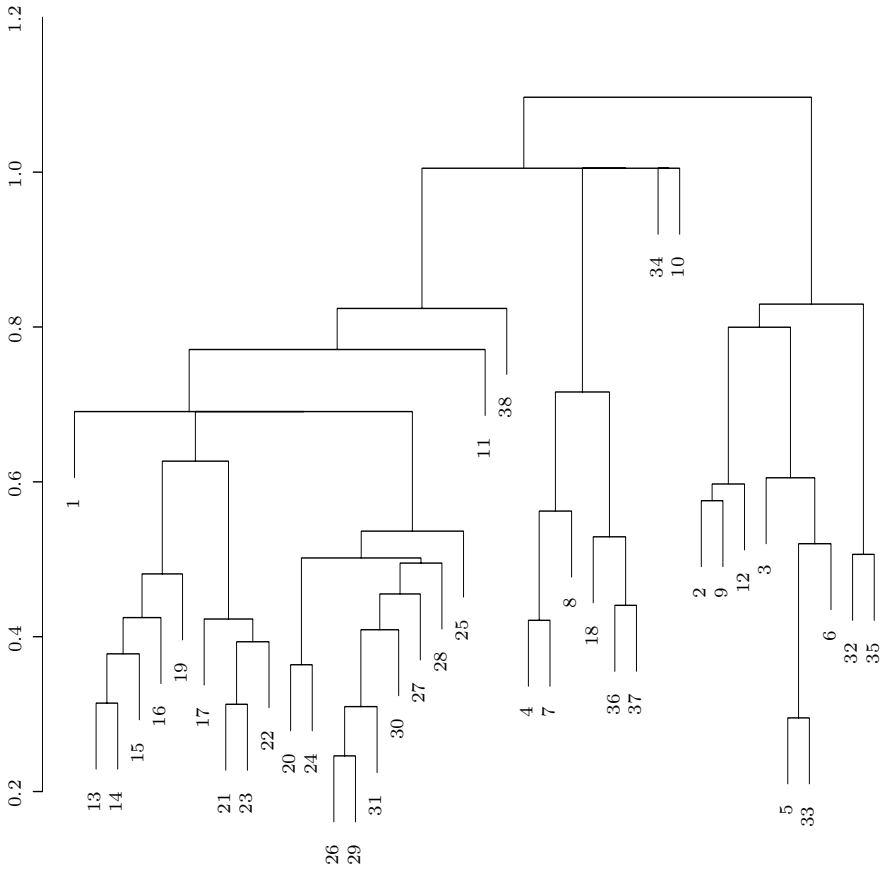

Figure 6. Hierarchical Clustering Dendrogram of Melanoma and Control samples. Divisive Clustering with correlation-based distance.

We again considered partitioning methods K-means and PAM, this time with K fixed at 3, thinking that the controls should cluster as a third group. Table 3

shows the number of samples assigned to combinations of clusters across methods (the controls are referred to as Bittner cluster 3 here). Only combinations which occurred (i.e., have nonzero entries) are given.

Agreement on cluster assignment of the controls is not especially high, perhaps due in part to heterogeneity in tissue type or biopsy site. However, the conclusion of Bittner et al. (2000) would seem to be that molecular profiling can distinguish subtypes of melanoma better than it can distinguish between melanoma or no melanoma. This situation suggests that the faith placed in the results of clustering algorithms may be misplaced.

Table 3. Cross-tabulation of cluster assignments.

Bittner	K-means	PAM	Number of Samples
1	1	1	11
1	1	2	2
1	2	2	6
2	1	1	1
2	2	2	4
2	2	3	1
2	3	1	2
2	3	3	4
3	2	3	1
3	3	1	3
3	3	3	3

4.3. Cluster validation methods

Bittner et al. (2000) used a variety of methods to assess the reliability of the cluster of 19 melanomas. A related problem is to determine the number of "true" clusters K in the data. This is a difficult problem in the field of clustering (Jain and Dubes (1988)), but one for which a number of solutions have been proposed.

Milligan and Cooper (1985) performed an extensive numerical comparison of 30 different methods. One of the methods they found to work well was that of Calinski and Harabasz (1974). In this method, the ratio of the between and within cluster sum of squares, appropriately normalized, is computed for each number of clusters K. The value of K that maximizes this ratio is taken to be the estimate of the number of clusters in the data. Another method was proposed by Krzanowski and Lai (1985). Their suggestion is to choose K to maximize a difference quantity based on the within clusters sums of squares. Neither of these methods allow for the case $K = 1$ (i.e., no clusters). A method that can test for the case of no clusters was developed by Hartigan (1975).

We applied these methods to the average linkage hierarchical clustering summarized in Figure 1. The number of clusters estimated by the Carabasz and Halinski (1974) method is $K = 2$. Based on this method, there is a cluster of 27 melanomas, not 19 as concluded by Bittner et al. (2000). Using the method of Krzanowski and Lai (1985) yields an estimate of eight clusters and does give the 19 melanoma cluster found by Bittner et al. (2000). Finally the method of Hartigan (1975) yields 2 clusters, so the melanoma cluster contains 27 samples.

When inferred clusters are supposed to represent some basic underlying biological process, the number of clusters represents an extremely important quantity. Yet these results cast additional doubt on the reliability of the 19 melanoma cluster found by Bittner et al. (2000), and highlight the difficulties inherent in assessing cluster reliability/reproducibility, and also of determining the number of clusters, a basic problem in cluster analysis.

4.4. Threshold level

Before performing the clustering, there was some preprocessing of the data. Values of the red/green intensity ratio that were lower than 0.02 were thresholded to 0.02, while those greater than 50 were thresholded to 50. Because these cutoffs, used by Bittner et al. (2000), seem somewhat arbitrary and without any apparent biological or data-driven justification, we considered alternative thresholds to see what, if any, effect the threshold level had on results. We looked at the following pairs of lower and upper threshold levels: (0.1,10), (0.3,10), (0.4,10), (1,10), and (1,3). Results did not vary much for more extreme thresholding of the data and, in fact, seemed to be more sensitive to choice of clustering method than thresholding. For a given metric, the results did not vary for more extreme thresholding of the data. However, the cutoff value on the dendrogram for the cluster of 19 melanomas increased with greater thresholding.

5. Identification of Genes Influencing Survival

As previously noted, the melanoma data set contains data on survival for 15 of the patients, 10 "clustered" (samples 13, 14, 15, 17, 18, 25, 28, 29, 30, 31) and 5 "unclustered" (samples 1, 7, 9, 10, 12). These data were used indirectly by Bittner et al. (2000) in an effort to identify genes which might influence survival. Because survival appeared to be higher in the clustered group (Figure 9), genes discriminating between the two groups perhaps played a role. We illustrate here how the survival data might be used directly in identification of genes potentially influencing survival, without performing any clustering. For further background on survival analysis, see, for example, Cox and Oakes (1984).

For individual i and gene $j = 1, \ldots, p$, we model the instantaneous failure rate, or hazard function $h(t)$, as a function of the expression level x_{ij} with the

Cox proportional hazard model $h(t) = h_0(t)exp(\beta x_{ij})$, so that the hazard functions of two different individuals are constant multiples of each other. Separate Cox models were fit for each individual gene. For each of the p genes, then, there is an estimated coefficient $\hat{\beta}$. These estimated coefficients can also be standardized as $t = \hat{\beta}/\hat{SE(\beta)}$. The pair of values $(\hat{\beta}, t)$ can be plotted for a visual representation of the values. Due to the tradeoff between statistical significance and practical importance, genes need to be prioritized according to some combination of these two criteria. Of highest interest are genes which have both large (in magnitude) standardized effect sizes (reflecting statistical significance) and also large (in magnitude) estimated effect sizes (reflecting practical importance, or biological meaning).

Standardized Cox Regression Coefficient vs. Coefficient

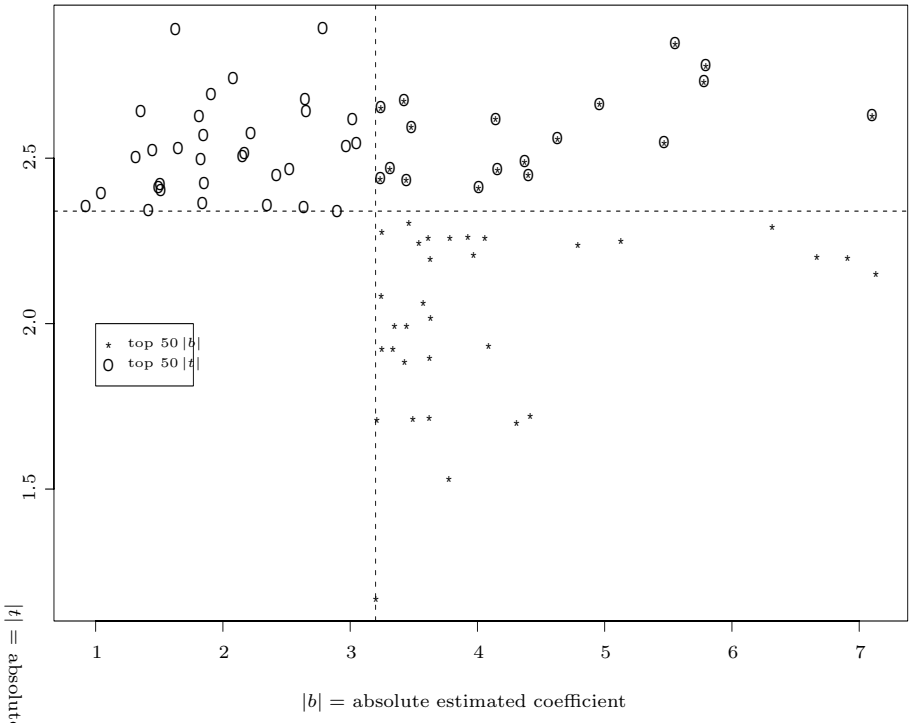

Figure 7. Identification of genes potentially influencing survival

Figure 7 plots $|t|$ versus $|\hat{\beta}|$ for genes in the top 50 in either estimate. The 50 cutoff is chosen for illustration. Cutoffs could be set higher or lower depending on the particular study, or chosen on other bases (e.g., $|\hat{\beta}|$ and $|t|$ exceeding some threshold of interest). The intersection of these two sets of genes occurs in the

upper right part of the plot (18 genes plotted by both plotting symbols). Of highest interest would be the five genes farthest to the upper right, and listed in Table 4.

Table 4. Sites potentially influencing survival.

Image clone ID	UniGene Cluster	UniGene Cluster Title
137209	Hs.126076	Glutamate receptor interacting protein
240367	Hs.57419	transcriptional repressor
838568	Hs.74649	cytochrome c oxidase subunit Vlc
825470	Hs.247165	ESTs, Highly similar to topoisomerase
841501	Hs.77665	KIAA0102 gene product

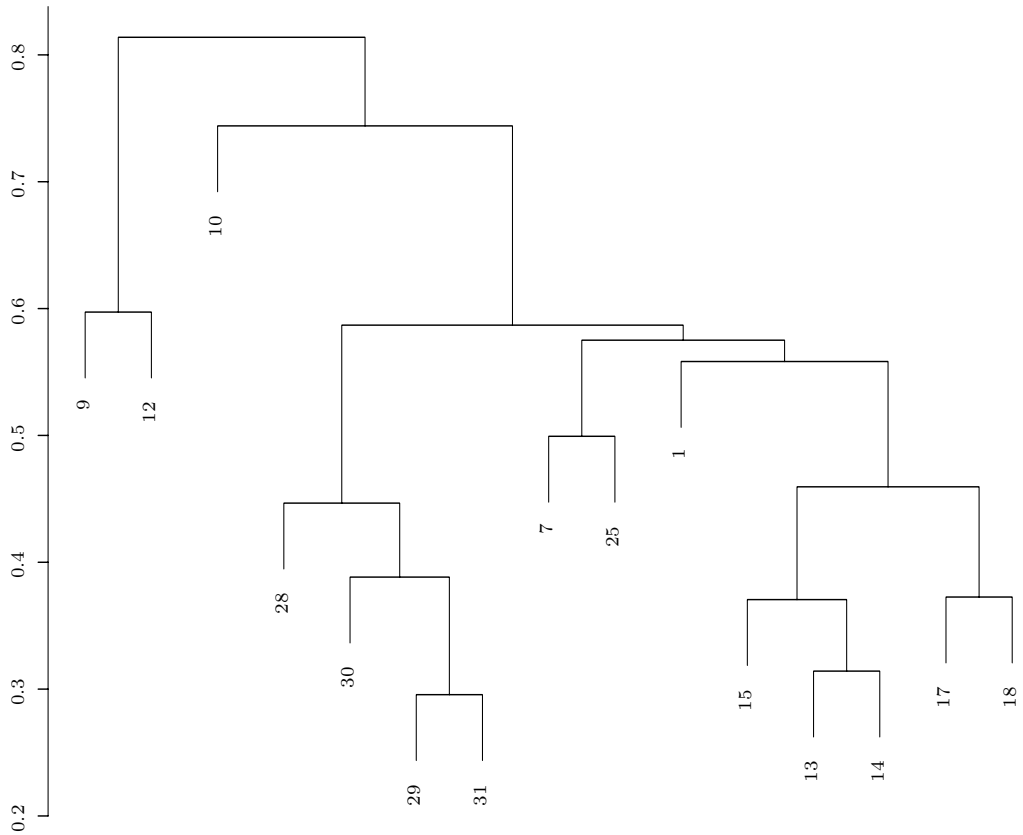

Figure 8. Hierarchical Clustering Dendrogram of Melanoma samples with Survival Data. Average linkage with correlation-based distance.

Somewhat surprisingly, none of these genes occurs in the weighted gene list of Bittner et al. (2000). One possible explanation is that the data set used for the identification of survival genes is only about half of the original set. Alternatively, it could be that the weighted scheme relies heavily on the cluster assignments of the 12 and 19 samples. In the reduced data set, these cluster assignments change. Figure 8 shows the dendrogram for average linkage hierarchical clustering (1 - correlation dissimilarity) of the 15 samples for which survival data are available. The original cluster assignment places samples 13, 14, 15, 17, 18, 25, 28, 29, 30, and 31 in cluster 1 and samples 1, 7, 9, 10, and 12 in cluster 2. It is clearly seen in the figure that samples 1 and 7 are now part of cluster 1.

If the new, reduced-data cluster assignment is used in the Kaplan-Meier analysis, the difference between the two survival curves becomes much smaller (Figures 9 and 10), with the p-value increasing from 0.135 to 0.472. The value of the clustering seems much less promising.

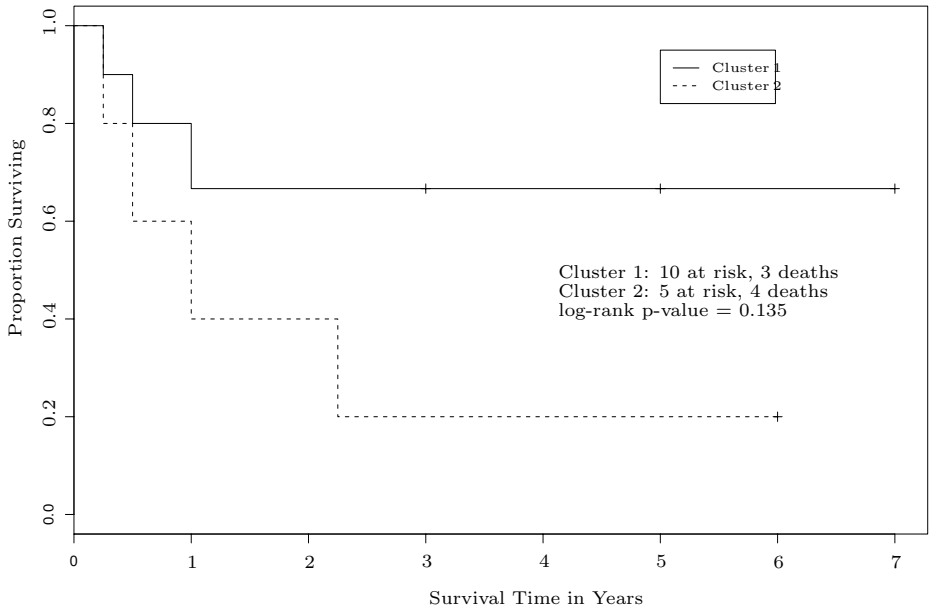

Figure 9. Kaplan-Meier survival plot: Bittner cluster assignment.

We have not attempted to carefully model the hazard function with combinations of genes, or to control the false positive rate in the presence of thousands of tests. The data may only be sufficient for suggestions of genes worthy of further investigation. With more extensive data, more careful modeling could be done.

Here, though, we do show an alternative strategy to clustering that ought to be useful for answering some of the types of questions that microarray experiments attempt to address.

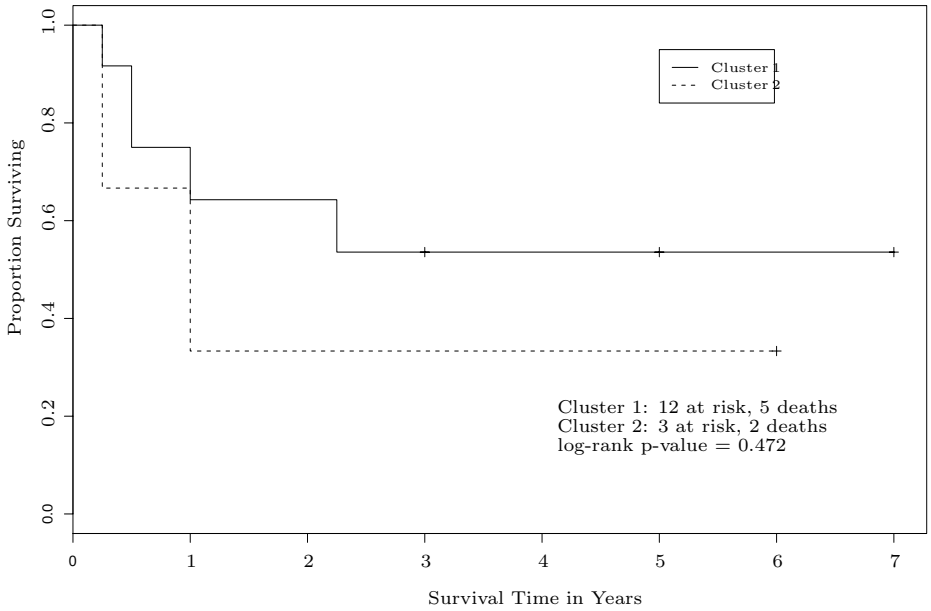

Figure 10. Kaplan-Meier survival plot: reduced data cluster assignment.

6. Discussion

We have illustrated here some of the problems which can occur in any data set when clustering samples of gene expression profiles. These include a possible high degree of dependence of results on choice of clustering algorithm, further dependence of results on the choice of samples to be included in the clustering (for example, whether or not to include control samples), and difficulty in assessing the validity of the grouping. With these problems inherent in the technique of cluster analysis, it is difficult to see how one might be capable of inferring a biological basis for the observed subgrouping. If the groupings are very tight, cluster analysis might then be appropriate and also more successful at suggesting true differences between groups. That was not the case for this melanoma data set.

Acknowledgements

We are most grateful to Sandrine Dudoit, Jane Fridlyand, and Terry Speed

of the University of California, Berkeley, for many helpful discussions and ideas. Two anonymous referees provided additional useful comments.

We would also like to express our appreciation to the Institute for Pure and Applied Mathematics (IPAM) and the National Science Foundation for support for participation in the program in Functional Genomics, Fall 2000, during which this work was initiated. This work was also partially supported by NIH grants 1 R01 GM59506 (DRG) and AG00057 (EMC).

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. and Staudt L. M. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**, 6745-6750.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics* **49**, 803-821.

Ben-Dor, A. Shamir, R. and Yakhini, Z. (1999). Clustering gene expression patterns. *J. Comput. Biology* **6**, 281-297.

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., Sondak V., Hayward N. and Trent J. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536-540.

Calinski, R. B. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Comm. Statist.* **3**, 1-27.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

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. (1993). *Cluster Analysis*. Arnold, London.

Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D. and Lander E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.

Gordon, A. (1999). *Classification*. Chapman and Hall/CRC Press, London.

Hartigan, J. (1975). *Clustering Algorithms*. Wiley, New York.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliff, NJ.

Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: an Introduction to Cluster Analysis*. Wiley, New York.

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.

Lance, G. N. and Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* **9**, 373-380.

Lipshutz R. J., Fodor S. P., Gingeras T. R. and Lockhart D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics* **21**(Suppl.), 20-24.

Lockhart D. J., Dong H., Byrne M. C., Follettie M. T., Gallo M. V., Chee M. S., Mittmann M., Wang C., Kobayashi M., Horton H. and Brown E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* **14**, 1675-1680.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium* **1**, 281-297.

Milligan G. W. and Cooper M. C. (1985). An examincation of procedures for determining the number of clusters in a data set. *Psychometrika* **50**, 159-179.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* **96**, 9212-9217.

Schena, M. (1999), editor. *DNA Microarrays: a Practical Approach*. Oxford University Press, Oxford.

Schena, M., Shalon, D., Davis, R. W. and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470.

Spellman P. T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D. and Futcher B. (1998). *Molecular Biology of the Cell* **9**, 3273-3297.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Amer. Statit. Assoc.* **58**, 236-244.

Department of Statistics, University of California, Los Angeles, California, U.S.A.
E-mail: darlene@stat.ucla.edu

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.
E-mail: ghoshd@umich.edu

Department of Statistics, Harvard University, Cambridge, Massachusetts, U.S.A.
E-mail: conlon@hustat.harvard.edu

(Received March 2001; accepted October 2001)