Exploratory screening of genes and clusters from microarray experiments

Robert Tibshirani[*], Trevor Hastie[†],
Balasubramanian Narasimhan[‡], Michael Eisen[§],
Gavin Sherlock [¶], Pat Brown[‖]
and David Botstein[**]

Abstract

We discuss a method called "cluster scoring" for supervised learning from a set of gene expression experiments. Cluster scoring generalizes methods that rank individual genes based on their correlation with an outcome measure. It begins with a clustering of the genes, for example from hierarchical clustering, and then computes outcome scores both for individual genes and the average gene expression for each of the clusters. A permutation method is used to identify the significant subset of these scores. We illustrate the method on both simulated data, and data from a study of lymphoma.

[*]Depts. of Health, Research & Policy, and Statistics, Stanford Univ, tibs@stat.stanford.edu

[†]Depts. of Statistics, and Health, Research & Policy, Sequoia Hall, Stanford Univ., CA 94305. hastie@stat.stanford.edu

[‡]Depts. of Statistics, and Health, Research & Policy, Sequoia Hall, Stanford Univ., CA 94305. naras@stat.stanford.edu

[§]Life Sciences Division, Lawrence Orlando Berkeley National Labs & Dept. of Molecular. and Cell Biology, University of California. Berk.; eisen@genome.stanford.edu;

[¶]Department of Genetics, Stanford University;botstein@genome.stanford.edu

[‖]Department of Biochemistry and HHMI, Stanford University;pbrown@cmgm.stanford.edu

[**]Department of Genetics, Stanford University;botstein@genome.stanford.edu

1 Background

In this paper we present a method for "supervised learning" from gene expression data. As our starting point, we have gene expression data from a collection of experimental samples, each hybridized to a micro-array— either cDNA or oligo arrays. Brown & Botstein (1999) and Dudoit et al. (2000) contain useful background information on microarrays.

We denote the expression values by x_{ij} for genes $i = 1, 2, \ldots p$ and samples $j = 1, 2, \ldots n$. Here typically $p >> n$. For oligo arrays these are the usual estimated expression levels; for cDNA arrays, x_{ij} is the log red/green ratio. We also have available a response measure $y = (y_1, y_2, \ldots y_n)$ for each sample (each y_j may be vector-valued). For example y_j might be a censored survival time, or a cancer class.

A number of methods have been proposed for evaluating the relationship between individual genes and the response y. When y takes on just two values, one can compute a standard t-statistic for each gene to assess its change over the two conditions. However with so many genes under assessment, control of the false positive rate can be important. Proposals along these lines include Tusher et al. (2001) and Dudoit et al. (2000). Tusher et al. (2001) also discuss other response types, including multiclass and survival data, Bayesian approaches were proposed by Efron et al. (2000) and Newton et al. (2000).

One shortcoming of these proposals is the fact that they operate on individual genes. Many genes are likely to operate in pathways, and hence will show significant correlation in expression. Clustering methods (see for example Eisen et al. (1998)) identify groups of genes with high correlation. In this paper, we use clustering methods as the basis for identifying both *individual genes*, and *clusters of genes*, that show significant correlation with a response measure.

2 A review of the SAM procedure

The basic proposal of this paper ("cluster scoring") is a generalization of the SAM (Significance analysis of microarrays) procedure of Tusher et al. (2001) for evaluating individual genes. We briefly review SAM, before describing cluster scoring.

For each gene i we define a score

$$d_i = \frac{r_i}{s_i + s_0}. \tag{1}$$

The quantity r_i is a measure of the relationship between the expression measurements for gene i and the response: for example, if y takes on just two values 1 and 2, then $r_i = \bar{x}_{i2} - \bar{x}_{i1}$, the difference in average expression values. For other response types, d_i is a score statistic, derived from an appropriate model. Tusher et al. (2001) discuss scores for many different types of re-

sponse measures, including one-class, two-class, multi-class, paired, survival and quantitative responses.

The quantity s_i is a standard deviation for gene i, and $s_0 > 0$ is an adjustment factor that prevents genes with very low expression levels from dominating the results.

Given the set of d_i values, $i = 1, 2, \ldots p$, our task is to decide which ones are "significantly" large. The SAM procedure is a way of thresholding the set of d_i values, and gives an estimate of the false positive rate of the resulting rule. A different but related proposal is given in Dudoit et al. (2000). The SAM procedure is outlined below.

The SAM procedure

1. Compute the order statistics of the d_i's: $d_{(1)} \cdots \leq d_{(p)}$

2. Take B sets of permutations of the response values y_j. [typically B=100 or 200 is sufficient]. For each permutation b compute statistics d_i^{*b} and corresponding order statistics $d_{(1)}^{*b} \leq d_{(2)}^{*b} \cdots \leq d_{(p)}^{*b}$. From the set of B permutations, estimate the expected order statistics by $\bar{d}_{(i)} = (1/B) \sum_b d_{(i)}^{*b}$ for $i = 1, 2, \ldots p$.

3. Plot the $d_{(i)}$ values versus the $\bar{d}_{(i)}$. For a fixed threshold Δ, starting at the origin, and moving up to the right, find the first $i = i_1$ such that $d_{(i)} - \bar{d}_{(i)} > \Delta$. All genes past i_1 are called "significant positive". Similarly, start at origin, move down to left and find first $i = i_2$ such that $\bar{d}_{(i)} - d_{(i)} > \Delta$. All

genes past i_2 are called "significant negative". For each Δ define the upper cut-point $\text{cut}_{up}(\Delta)$ as the smallest d_i among the significant positive genes, and similarly define the lower cut-point $\text{cut}_{low}(\Delta)$.

4. For a grid of Δ values, compute the total number of significant genes (from the previous step), and the expected number of falsely called genes, by computing the proportion of values among each of the B sets of $d^{*b}_{(i)}$, $i = 1, 2, \ldots p$, that fall above $\text{cut}_{up}(\Delta)$ or below $\text{cut}_{low}(\Delta)$.

5. The user then picks a value for Δ and the significant genes are listed.

Figure 1 shows an example taken from Tusher et al. (2001). There are 7129 genes and 8 samples, 4 in each of two classes (untreated and treated). In the figure we have chosen $\Delta = 1.2$ (based on the false positive rate), giving the cut-points $\text{cut}_{low} = -3.3$, $\text{cut}_{up}(1.2) = 3.3$, and 46 genes are called significant. The expected number of false positive genes is about 9. There are 24 called genes in the top right corner (red), and 22 called genes in the bottom left (green). One advantage of the use of the Δ band to define cut-points, is the potential for asymmetry. That is, $\text{cut}_{up}(\Delta)$ may not equal $-\text{cut}_{low}(\Delta)$ (although they are nearly equal in Figure 3) and hence more positive (or negative) genes can be called significant.

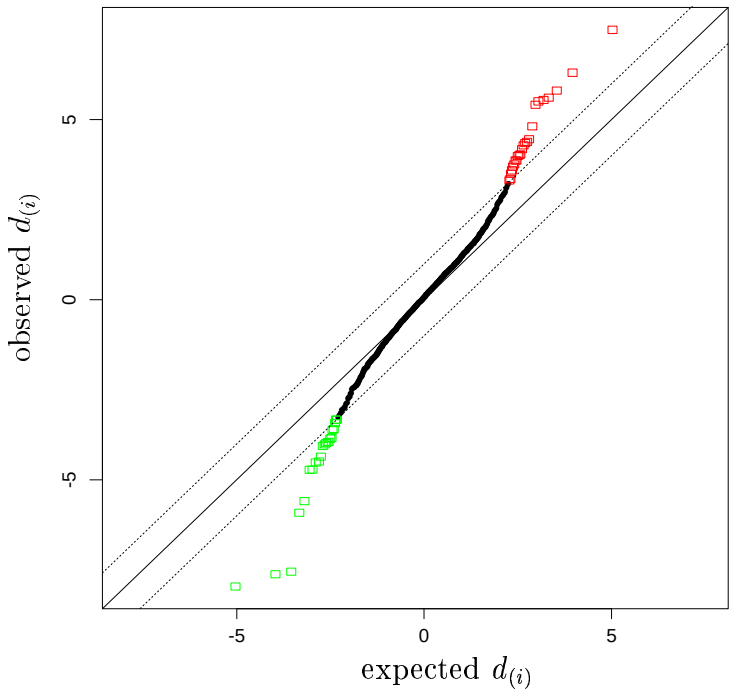

Figure 1: *SAM plot for some two-class data*

3 Cluster scoring

We now give the main proposal of this paper. We start with any clustering of the genes. In this paper we use (average linkage) hierarchical clustering, because of its popularity in micro-array analysis. We discuss the choice of clustering method in section 4.2 below.

For all clusters we compute the average gene expression: with hierarchical clustering, clusters from all levels of the dendrogram are used. With p genes, there are $p-1$ such clusters. Then we apply SAM to the collection of both individual genes and the cluster averages.

In some datasets, there might be a large cluster of highly correlated genes, each of which is highly correlated with the response. In that instance, the cluster's average will also be highly correlated with the response. In other datasets, a single gene might be highly correlated with the response. We'd like our procedure to work well in both situations.

Hence some special modification are needed to a) encourage larger clusters to emerge as significant (since they're in competition with a large number of individual genes) and b) thin out the large set of clusters to focus on the interesting ones. We give the *cluster scoring* procedure below, and then give details of the modifications.

The use of the quantity $s_0(p_c)$ in step (2) encourages larger and tighter clusters. It is an estimate of the standard deviation of d_i for cluster size p_c.

Cluster scoring

1. Start with a hierarchical clustering of the genes. Denote a cluster of genes by c, and the corresponding gene expression average by $\bar{x}_c = (\bar{x}_{c,1}, \bar{x}_{c,2}, \ldots \bar{x}_{c,n})$. There are $2p-1$ clusters, including individual genes. Let p_c be the number of genes in cluster c.

2. Define the score for each average gene expression $\bar{x}_c$ as

$$d_c = \frac{r_c}{s_c + s_0(p_c)} \tag{2}$$

   Here r_c measures correlation with the response, s_c is standard deviation of the cluster average $\bar{x}_c$, and $s_0(p_c)$ is an adjustment factor (details below).

3. For each gene i let $c(i)$ be the set of clusters containing it. Let

$$\hat{c}(i) = \arg\max_{c \in c(i)} |d_c| \tag{3}$$

   the *winning cluster* for gene i. Let R be the set of winning clusters.

4. Apply the SAM procedure to the clusters $c \in R$. When computing false positives, count unique genes.

In more detail, by including the factor $s_0(p_c)$:

- for a given cluster size, the procedure rewards clusters having highly correlated genes, and hence large values of s_c, and
- overall, it rewards larger clusters, since $s_0(p_c)$ will tend to be smaller for larger clusters.

As an example of the first point, suppose there are two clusters of the same size, the first being a tight cluster with $s_c = 10$, and the second being less tight, with $s_c = 1$. If $s_0(p_c) = 5$ say, then inclusion of this factor will reduce the value of d_c for the first cluster by a factor $10/(10 + 5) = 2/3$, and that for the second cluster by $1/(1 + 5) = 1/6$. Hence the score for the tighter cluster is reduced less, and thus is favored.

The quantity $s_0(p_c)$ is computed by dividing the range of cluster sizes into 100 quantiles $q_1 = 0, q_2, \ldots q_{100}$, and then setting $s_0(p_c)$ equal to the α quantile of $\{s_c; p_c \in (q_j, q_{j+1}]\}$. In this paper we set $\alpha = .5$.

Step 3 thins out the clusters, keeping only those that are the "winning" cluster for some gene.

Note that we count unique genes in computing the false positive rate: since the clusters are overlapping this avoids counting a given gene more than once,

4 An Example

We started with expression measurements of 3906 genes over 40 patient samples, from a study of diffuse large cell lymphoma (Alizadeh et al. (2000)). Diffuse large B-cell lymphoma (DLBCL), the most common subtype of non-Hodgkin's lymphoma, is clinically heterogeneous: 40% of patients respond well to current therapy and have prolonged survival, whereas the remainder succumb to the disease. The objective here is to relate gene expression to patient survival.

In order to test the procedure, we used the expression values but not the observed survival times. Instead we simulated an artificial set of survival times, so that the "true" cluster could be defined.

To carry this out, hierarchical clustering was applied to the genes, and then a moderate-sized cluster of 162 genes was selected at random from the resulting set of clusters. Denote this cluster by c_0. A set of standard Gaussian survival times for the 40 patients was simulated, having correlation .95 with the average gene expression of cluster c_0. (.95 being a high but plausible value in practice; later this is reduce to .75) Then we applied the cluster scoring technique to the data. The score d_i is based on Cox's score statistic for survival data, detailed in Tusher et al. (2001).

The results are shown in Figures 2, 3, and 4. Figure 2 shows the set R of 381 "winning" clusters from step 3 of the cluster scoring procedure.

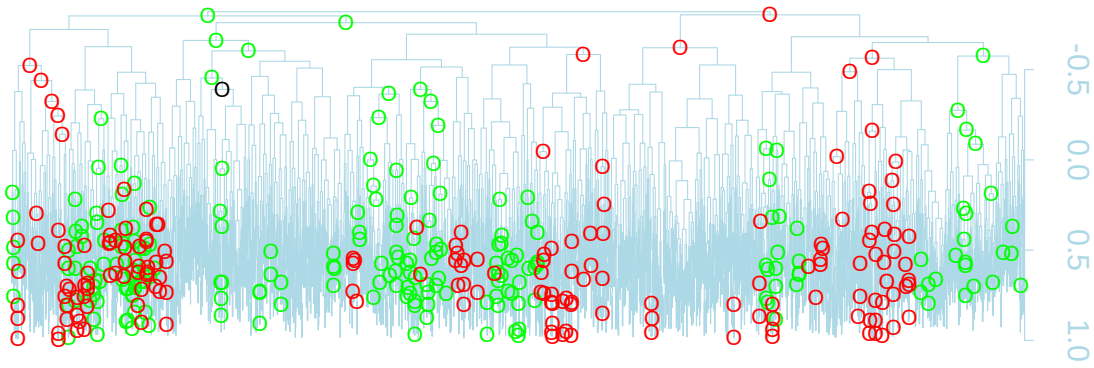

Figure 2: *Results for simulated data, correlation =.95. The true generating (negative) cluster is black. Shown is the set R of 381 "winning" clusters from step 3 of the cluster scoring procedure. Red clusters have a average gene expression that is positively correlated with the response, while green ones are negatively correlated with the response. The true black cluster also belongs to R. The correlation of the average gene expression from each cluster with the response ranged from $-.77$ to* .95.

The SAM plot in the middle right panel of Figure 3 calls 7 clusters significant: these are marked on the dendrogram of Figure 4. The other panels in Figure 3 are explained in the figure caption.

We see in Figure 4 that the chosen clusters are "sons" of the preassigned true cluster (marked in black). There are 162 unique genes in the called clusters, with zero false positive genes. Hence the procedure identified the correct set of genes exactly.

Figure 5 shows the sizes of the top 100 scoring clusters, as the factor s_0 is varied from the 0th percentile to the 100 percentile of the s_i values. The cluster sizes increase slightly, with a marked increase at the 100th percentile. In our example we use the 50th percentile, as a reasonable intermediate value.

Figure 6 compares cluster scoring to the SAM procedure for screening individual genes, as the threshold for each method is varied. We see that cluster scoring, by explicitly looking for clusters, has fewer false positive genes for the same number of genes called significant.

Figure 7 shows the result when the correlation between the survival times and the average gene expression of the "true" cluster is reduced from .95 to .75. The results degrade a little, as there are now 3 false positive clusters.

How large is a realistic correlation in this setting? With so many genes, moderately large correlations with any response variable can be expected

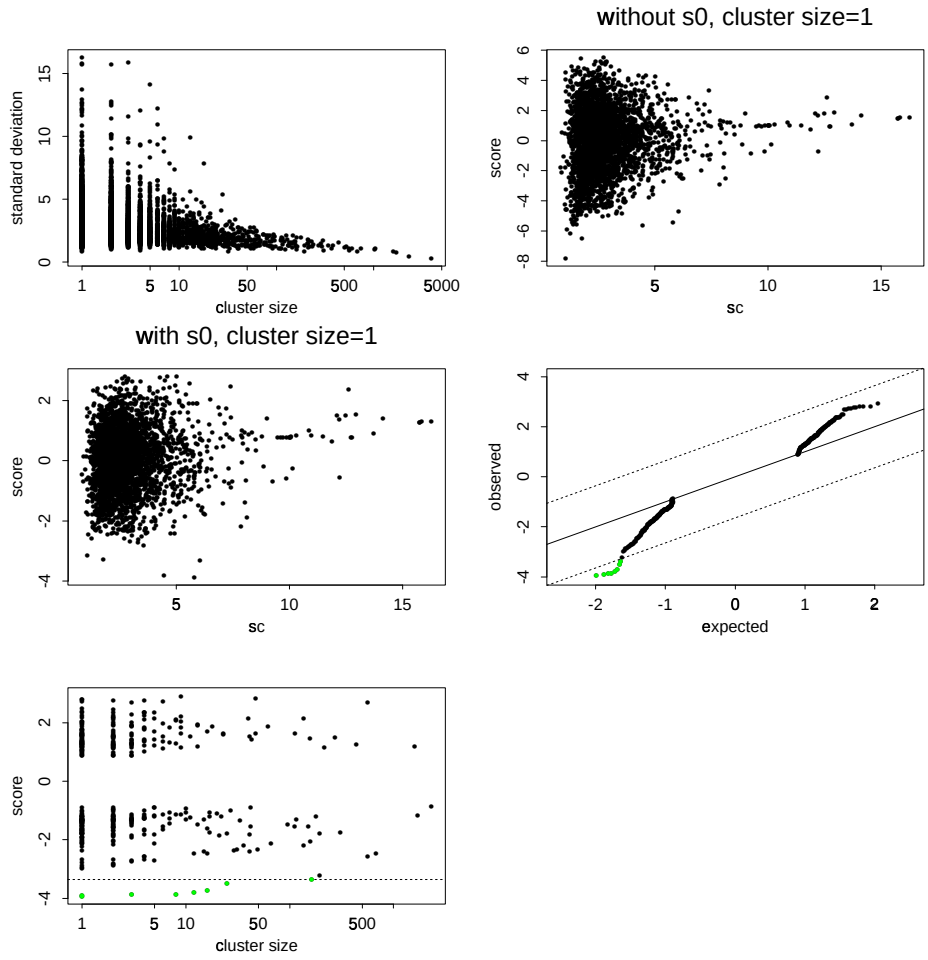

Figure 3: *Results for simulated data, correlation =.95. Top left panel shows how the standard deviation s_c tends to decrease with cluster size. The top right panel shows the score d_c as a function of the standard deviation s_c, without the factor s_0: smaller clusters have the largest (negative) scores. s_0 is used in the middle left panel, and now some clusters with larger standard deviations are favored. The middle right panel shows the SAM plot with $\Delta = 1.64$. The observed scores d_c are plotted on the vertical axis, versus the expected scores under permutation of the responses values. Finally the bottom left panel displays the SAM cutoff for the score d_c, and the resulting chosen clusters below the broken line.*

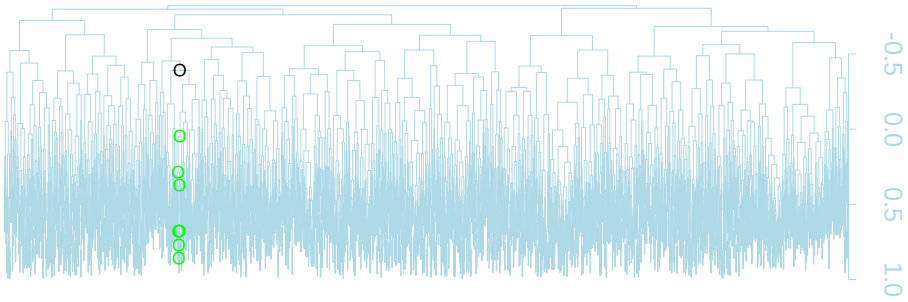

Figure 4: *Results for simulated data, correlation =.95. The true generating (negative) cluster is black and was called by the procedure, as were the green clusters.*

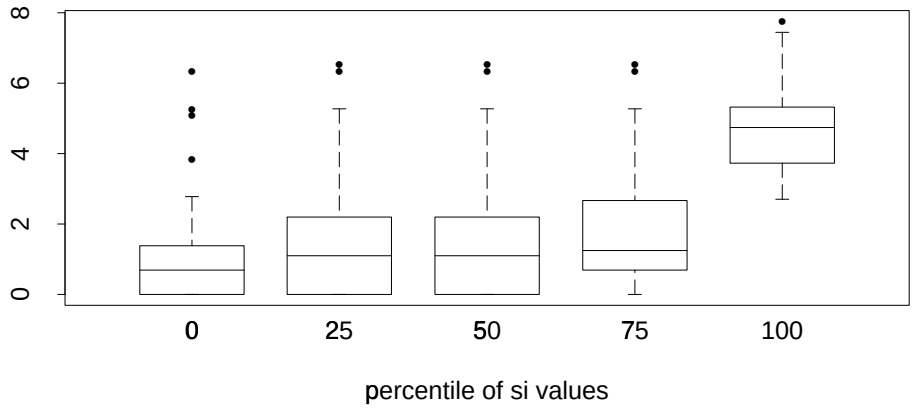

Figure 5: *Sizes of the top 100 scoring clusters, as the factor s_0 is varied from the 0th percentile to the 100 percentile of the s_i values.*

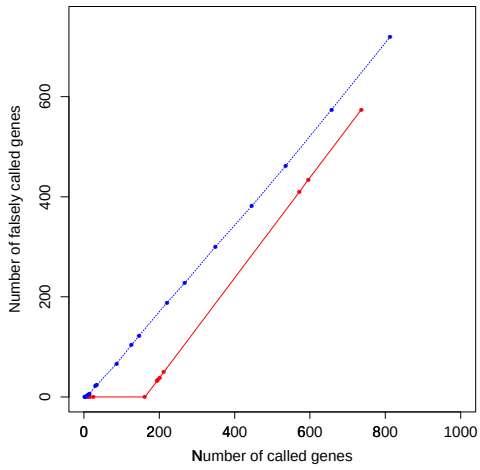

Figure 6: *Comparison of cluster scoring to individual gene screening via SAM. Shown are number of falsely called genes versus number of called genes, for SAM (blue, broken) and cluster scoring (red, solid).*

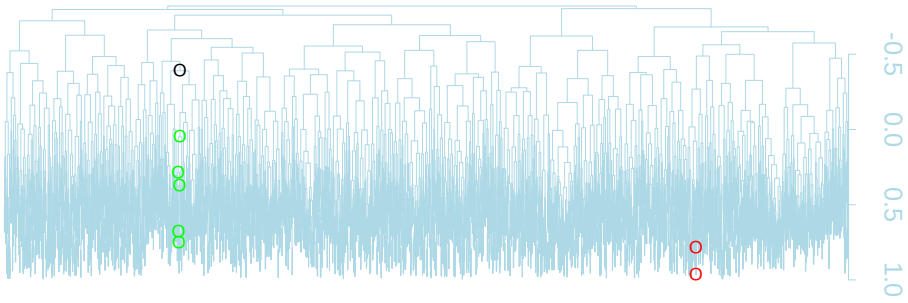

Figure 7: *More results for simulated data, correlation=.75. The true generating (negative) cluster is black and was called by the procedure, as were the green clusters. The red clusters are the positive called clusters.*

to occur just by chance. To quantify this, we fixed the expression data and generated 50 sets of survival times as independent standard Gaussian variates. The average maximum (absolute) correlation of the survival times with gene expression was 0.54. Hence correlations of .75 and above are not much over "noise" level in this problem. Note that this example was based on a single simulation. A broader study would be useful in understanding the properties of the method. We study one such property next.

4.1 Simulations

We generated 50 replications from the scenario described above. The objective was to investigate how the cluster sizes chosen by the method compared to the "true" cluster size, and to measure the correlation of the chosen cluster with that of the "true" cluster.

The cluster size of the "true" cluster was chosen at random from the set of unique cluster sizes, and then vector of response values y were generated by adding standard Gaussian noise to the gene expression average of this true cluster. Fifty permutations were used in the SAM procedure. The standard deviation of the noise was chosen so that the correlation of y with the gene expression average of this true cluster had mean about .75 across the 50 simulations. We then applied cluster scoring to each of the 50 datasets, giving the results in Figure 8. The left panel shows the cluster size of the cluster with highest score, versus the size of the true cluster. The procedure sometimes underestimates the true cluster size but overall does quite well. The right panel plots the correlation of the gene expression average of the highest score cluster and the true cluster. This correlation averages around .75, which is to be expected from the noise level in the data.

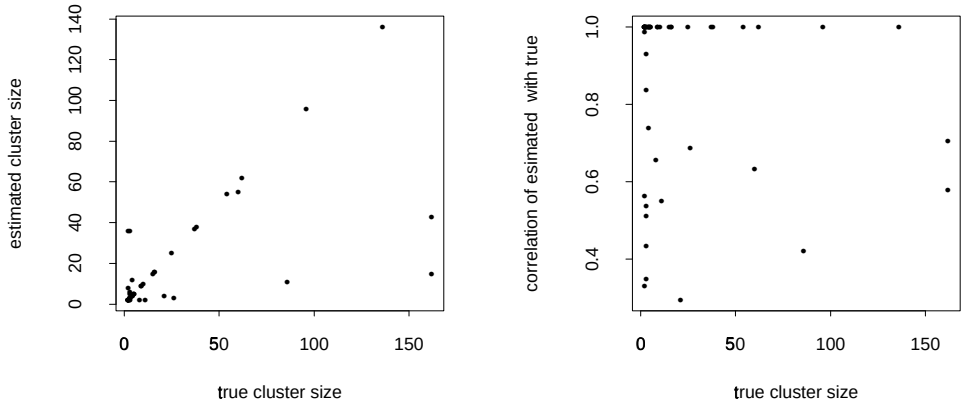

Figure 8: *Results from 50 simulations. Left panel shows the cluster size of the cluster with highest score, versus the size of the true cluster. The right panel plots the correlation between the average gene expression of the highest score cluster and that for the true cluster, versus the true cluster size.*

4.2 Choice of clustering method

In this paper we use hierarchical clustering of the genes, as input into the cluster scoring procedure. However any set of gene clusters could be used. Possibilities include:

- k-means clustering
- self-organizing maps (Kohonen 1990)
- gene shaving (Hastie et al. 2000)
- plaid models (Lazzeroni & Owen 2000)

Some of these methods (gene shaving, plaid models) allow genes to appear in more than one cluster. To make the cluster scoring procedure most effective, it is best to have a wide range of cluster sizes in the available set. Hierarchical clustering provides such a set; one could use k-means clustering for many different values of k, to achieve a similar range of cluster sizes. We have not yet experimented with cluster scoring, using other clustering methods.

5 Lymphoma data

In this section we analyze the lymphoma data mentioned earlier, here using the actual survival times rather than simulated ones.

Table 1: *Lymphoma data: number of called genes, and estimated number of false positives for various values of the tuning parameter* Δ.

Δ	Aver # false calls	# called	False discovery rate
0.0	2713.25	3258	0.83
0.1	1290.00	1831	0.70
0.2	71.35	156	0.46
0.3	11.45	33	0.35
0.4	0.00	0	-

The left panel of Figure 9 shows the observed versus expected plot for the cluster scores. With a value of $\Delta = .3$, the resulting cutoff was -1.78, and gave the clusters shown in Figure 10. They are small– ranging in size from 1 to 4 genes, 33 unique genes in all.

Table 1 shows the number of called genes, and estimated number of false positives for various values of the tuning parameter Δ. For $\Delta = .3$, about one third of the 33 called genes are expected to be false positives. This high false discovery false is about the same as that obtained from SAM, and indicates that the correlation between gene expression and survival time is not great.

The significant gene clusters may be worthy of further investigation. They are shown in Figure 10. Many of the genes are ESTs (expressed sequence tags). A nice feature of the Cluster Scoring procedure is that each gene is

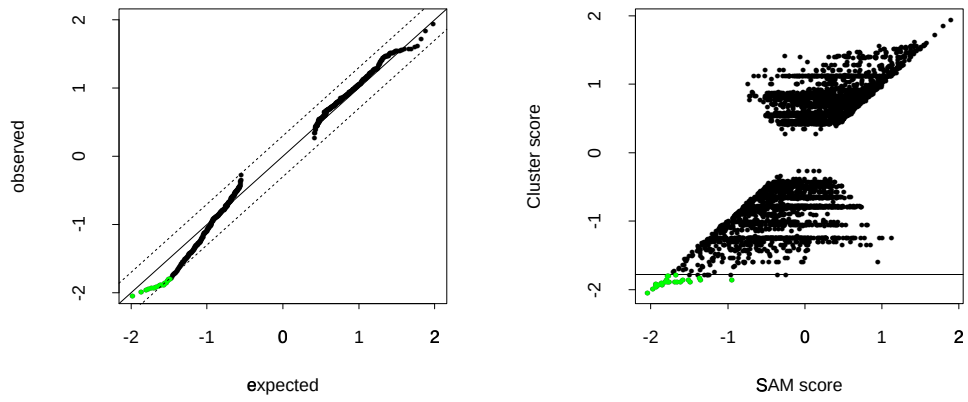

Figure 9: *Left panel: Observed vs expected plot for actual lymphoma data. Right panel: Cluster score versus SAM score, for each gene.*

assigned a score, the largest score from any cluster in which it appears. The right panel shows these cluster scores versus the SAM scores, for each gene. We see that many of the genes called significant (below the broken line) have individual scores which are not very large (≈ -1).

In general, the results in this example are somewhat disappointing. No large significant clusters were found. This does not seem to be a fault of the cluster scoring method, but is due to a general lack of correlation between expression and outcome in these data. Our collaborators are currently;collecting a larger set of samples, and these may provide more useful results.

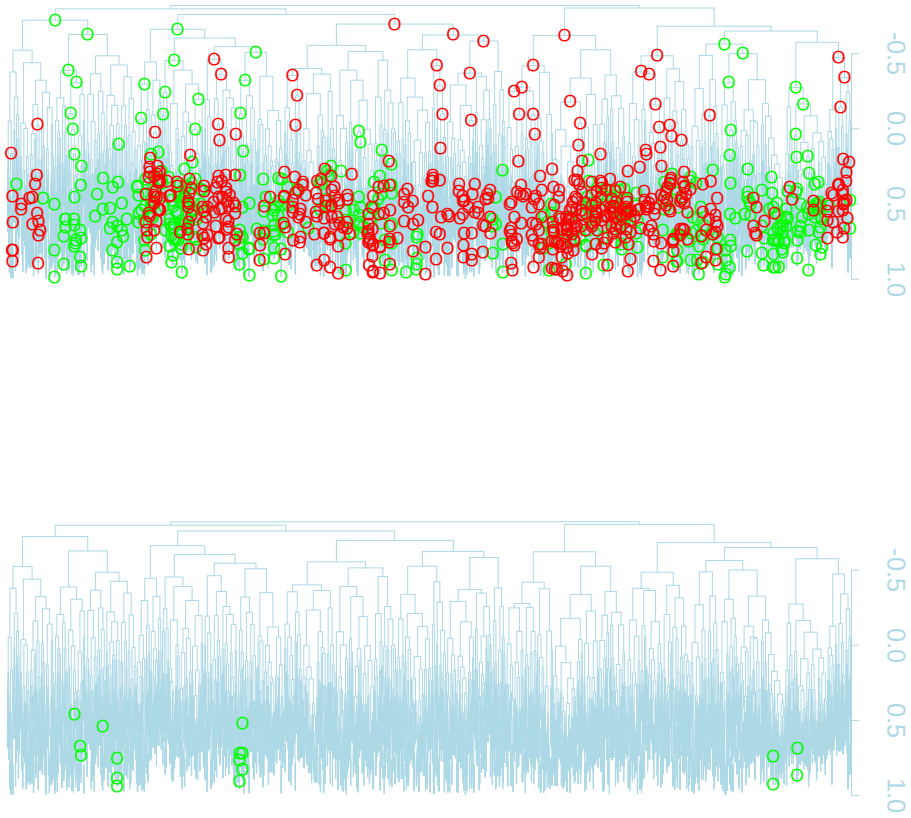

Figure 10: *Winning clusters (top) and significant clusters (bottom) for the actual lymphoma data*

6 Discussion

Cluster scoring is a promising method for finding both individual genes and clusters that have significant correlation with a response measure. It can be used for a wide variety of outcome measures, including survival categorical and paired data.

Associated with the method is an estimate of the false discovery rate, counting the number of unique false positive genes in the clusters that are called significant. However the development in this paper is exploratory, and we make no rigorous claims about the statistical properties of our method. Further investigation of these properties is needed, including study of type I error rates, power and false discovery rate.

A new version of the SAM software may soon be available, implementing the cluster scoring methodology, and interfacing with both Xcluster [1] and Cluster and TreeView[2].

Acknowledgments: Tibshirani was partially supported by NIH grant 2 R01 CA72028, and NSF grant DMS-9971405. Hastie was partially supported by grant DMS-9803645 from NSF and grant ROI-CA-72028-01 from the National Institutes of Health. The authors would like to thank two referees for helpful comments.

[1] http://genome-www.stanford.edu/~sherlock/

[2] http://rana.lbl.gov/

References

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., J., P., Marti, G., Moore, T., Hudsom, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R. Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000), 'Identification of molecularly and clinically distinct substypes of diffuse large b cell lymphoma by gene expression profiling', *Nature* **403**, 503–511.

Brown, P. & Botstein, D. (1999), 'Exploring the new world of the genome with dna microarrays', *Nature genetics* pp. 33–37.

Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2000), Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Unpublished, available at http://www.stat.berkeley.edu/users/sandrine.

Efron, B., Tibshirani, R., Goss, V. & Chu, C. (2000), Microarrays and their use in a comparative experiment. submitted.

Eisen, M., Spellman, P., Brown, P. & Botstein, D. (1998), 'Cluster analysis and display of genome-wide expression patterns', *Proc. Nat. Acad. Sci* **95**, 14863–14868.

Hastie, T., Tibshirani, R., Eisen, M., Alizadeh, A., Levy, R., Staudt, L., Botstein, D. & Brown, P. (2000), 'Identifying distinct sets of genes with similar expression patterns via "gene shaving"', *Genome Biology* **1(2)**, 1–21.

Kohonen, T. (1990), 'The self-organizing map', *Proc. of IEEE* **78**, 1464–1479.

Lazzeroni, L. & Owen, A. (2000), Plaid models for gene expression data, Technical report, Stanford University. Available at "http://www-stat.stanford.edu/ owen.

Newton, M., Kendziorski, C., Richmond, C., Blatter, F. & Tsui, K. (2000), On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. To appear, J. Comp. Biology.

Tusher, V., Tibshirani, R. & Chu, C. (2001), 'Significance analysis of microarrays applied to transcriptional responses to ionizing radiation', *PNAS* pp. 5116–5121.