Pacific Symposium on Biocomputing 6:52-63 (2001)

A NONPARAMETRIC SCORING ALGORITHM FOR IDENTIFYING INFORMATIVE GENES FROM MICROARRAY DATA

PETER J. PARK, MARCELLO PAGANO, and MARCO BONETTI
Department of Biostatistics, Harvard School of Public Health
655 Huntington Ave., Boston, MA 02115
{*ppark,pagano,bonetti*}*@hsph.harvard.edu*

Microarray data routinely contain gene expression levels of thousands of genes. In the context of medical diagnostics, an important problem is to find the genes that are correlated with given phenotypes. These genes may reveal insights to biological processes and may be used to predict the phenotypes of new samples. In most cases, while the gene expression levels are available for a large number of genes, only a small fraction of these genes may be informative in classification with statistical significance. We introduce a nonparametric scoring algorithm that assigns a score to each gene based on samples with known classes. Based on these scores, we can find a small set of genes which are informative of their class, and subsequent analysis can be carried out with this set. This procedure is robust to outliers and different normalization schemes, and immediately reduces the size of the data with little loss of information. We study the properties of this algorithm and apply it to the data set from cancer patients. We quantify the information in a given set of genes by comparing its distribution of the score statistics to a set of distributions generated by permutations that preserve the correlation structure among the genes.

1 Introduction

Simultaneous measurements of expression levels for thousands or even tens of thousands of genes have become feasible through the DNA microarray technology. These high-throughput methods measure the abundance of mRNAs transcribed during gene expression through the process of hybridization. The gene expression patterns in the microarray data have already provided some valuable insights in a variety of problems, and it is expected that knowledge gleaned from microarray data will contribute significantly to advances in fundamental questions in biology as well as in clinical medicine. [1–5,7,8,11–13]

While the most obvious patterns in expression levels can be detected even through visual inspection, there is much more information encoded in the data. Systematic and sophisticated methods for extracting all the significant information are necessary to take full advantage of the technology. For instance, it is common to discard the measurements when the fold difference is less than a certain threshold level, if that level is considered to be within the range of noise in the data. However, we know that many biological processes of interest

involve small fold changes below such cut-off levels. On the other hand, it may be tempting to assign significance to an expression pattern which may support a particular hypothesis, when there is a reasonable probability that the pattern may have been observed due to chance. Clearly, both the laboratory techniques and data analysis techniques need to be improved.

One application of microarray data is characterization of different cell types. Alon *et al.* (1999) analyzed a data set consisting of the gene expressions in 40 tumor and 22 normal colon tissue samples, and found coherent patterns and clustering in different families of genes.[2] Golub *et al.* (1999) developed methods for identifying new cancer classes and making prediction for the classes to which new tumors belong, thus demonstrating molecular classification of cancer by gene expression monitoring. Ross *et al.* (2000) described how gene expression profiles characterize patterns of phenotypic variation in the 60 cancer cell types.[12] Alizadeh *et al.* (2000) studied the diversity of gene expression in diffuse large B-cell lymphoma and were able to identify previously undetected and clinically significant subtypes of cancer.[1]

A problem that arises in this context is the following. What genes are useful for classification and how many of them should be used for predicting the classes of new samples? Identification of the informative genes is beneficial in that those genes may reveal insights into the biological process. It is also important to pre-process the data and reduce its size, as one of the major problems in studying the microarray data is the high dimensionality due to a large number of genes. A part of the data that contains no useful information should be excluded as soon as possible so that more detailed analysis, such as Self-Organizing Map (SOM) [10], hierarchical clustering [6], or construction of relevance network [4] can be carried out. In the SOM approach, this reduction step is particularly significant, in order to prevent nodes from being attracted to large sets of invariant genes.

To eliminate those genes that are likely to be insignificant, one can construct a filter, for example, a variation filter that excludes genes with less than five-fold variation across the collection of samples [7] or a variation filter with some threshold numbers for both relative and absolute changes.[14]

We propose a nonparametric scoring algorithm that is more systematic and robust. It is more refined than the ones previously described, yet is fast and simple to understand. Because this scheme uses only the ranks rather than the actual expression levels, there may be a slight loss of information. However, in return we gain a valid test with robustness to outliers, normalization schemes, and systematic errors such as chip-to-chip variation.

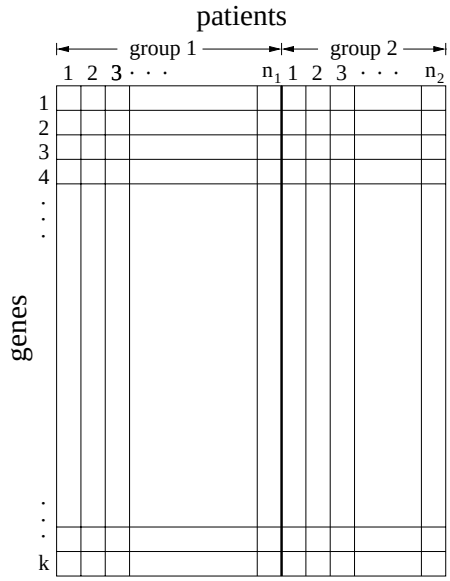

Figure 1: Microarray data have a large number of genes for few patients. Data are divided into two groups, with n_1 and n_2 patients in groups 1 and 2, respectively.

2 Method

Typical microarray data consist of expression levels for a large number of genes on a relatively small number of samples. For each gene, we have the expression levels for all the patients. We assume that the data contain no missing value; if there are missing values, we may use a variety of methods, for example, an EM-type algorithm[11], to impute the values. We assume that there are a total of n patients in two groups, with n_1 patients in the first group and n_2 in the second group. The same approach as presented below can be extended easily to the case of more than two groups.

2.1 Algorithm

We first sort the data so that the patients in the first group are on the left and those in the second group are on the right, as shown in Figure 1. Then, for each gene, we compute a score that captures the extent to which the gene is differentially expressed in the two groups. First, we assign 0's to the n_1 patients in group 1 and 1's to the n_2 patients in group 2. Then we sort the expression

levels from the smallest to the largest, at the same time permuting the 0's and 1's along with the expression levels so that the resulting sequence of 0's and 1's indicates the group membership of the patients having the ordered expression levels. How closely the 0's and 1's are grouped together is a measure of the correspondence between the expression levels and the group membership. If a particular gene can be used to divide the groups exactly, one would observe a sequence of all 0's followed by all 1's, or vice versa.

As an example, consider the following case. Suppose we have $n = 6, n_1 = 3$, and $n_2 = 3$. Then we have

```
Patients:            1    2   3   4   5   6

Expression Levels:  95  106  20  74  69 271
Groups:              0    0   0   1   1   1

After sorting:
Expression Levels:  20   69  74  95 106 271
Groups:              0    1   1   0   0   1
```

Based on this sequence, we can compute a score statistic that measures the disorder of 0's and 1's for each gene in the following way. We define the score to be the smallest number of swaps of consecutive digits necessary to arrive at a perfect splitting, with all the 0's on the left and all the 1's on the right. In the above example, we have the score of 4 as seen below.

score:	data:						positions swapped
	0	1	1	0	0	1	
+1	0	1	0	1	0	1	3 and 4
+1	0	0	1	1	0	1	2 and 3
+1	0	0	1	0	1	1	4 and 5
+1	0	0	0	1	1	1	3 and 4

This score can be shown to be an extreme case of the Kendall's τ [a] statistic in which the ranks are collapsed to two categories.[9]

[a] Kendall's Rank Correlation Coefficient τ is a nonparametric measure of association based on the number of concordances and discordances in paired observations. Concordance occurs when paired observations vary together, and discordance occurs when paired observations vary differently.

Then we can write

$$\text{Score} = \sum_{i \in \mathcal{N}_2} \sum_{j \in \mathcal{N}_1} h(x_j - x_i),$$

where $\mathcal{N}_i$ represents the set of indices belonging to group i and $h(x)$ is the indicator function

$$h(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1, & \text{if } x > 0. \end{cases}$$

This formulation can be interpreted as counting, for each element in the second group, the number of elements in the first group that are larger, and summing these numbers. (Equivalently, one can count the number of smaller elements in the second group compared to each of the first group.)

It is easy to notice that we have defined the score above in such a way that both a low and a high score indicate a differentially expressed gene. This was a result of moving the 0's to the left rather than allowing them to move to either side. We could have used the symmetry to define a score that is the minimum of the two possible arrangements, but we chose the former way so that we have a symmetric distribution of scores. The maximum score is $n_1 n_2$, which happens when the digits are completely in reverse. In that case, n_1 digits need to be moved n_2 spaces away, or vice versa, for the total of $n_1 n_2$ swaps.

2.2 Application

We apply this algorithm to the leukemia data analyzed by Golub *et al.* (1999).[7] [b] This data set contains 38 bone marrow samples of acute leukemia patients, belonging to two groups. There are 27 patients with acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML). High-density oligonucleotide microarrays containing 7129 probes for 6817 human genes were used.

When we run our algorithm on this data set, we find that there is in fact one gene, Zyxin, that has the score 0; that is, it splits the two classes exactly. Zyxin is known to encode proteins for cell adhesion. Next best scores are 2 (Cystatin C), 5 (Leukotriene C4 synthase and Elastatse 2), and 6 (Cystatin A and CD33 antigen).

In Figure 2, we have plotted on the left the profile of a gene, randomly selected among the genes with scores in the middle of the range. The first 27 patients on the x-axis belong to the ALL group and the remaining 11 to the AML group. Clearly, this gene makes no distinction between the two

[b] This data set is publicly available at http://www.genome.wi.mit.edu/MPR/

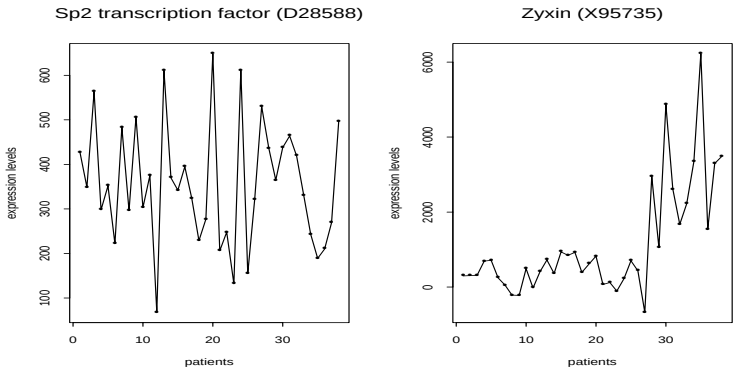

Figure 2: We plot the expression levels for two different genes. There are 38 patients, indicated on the x-axis, first 27 of which are in group 1 (ALL) and the remaining 11 are in group 2 (AML). Sp2 transcription factor on the left had a score close to half the maximum, hence among the least informative. Zyxin had a score of 0: its expression level is uniformly low for the first 27 patients and higher for the rest.

groups. On the right, we plotted the profile of Zyxin, which has uniformly low expression levels for the first group and higher levels for the second group.

If the gene had been completely independent of the classification, the chance of having a gene with score 0 or the maximum $n_1 n_2$ is $2\,n_1!\,n_2!/n! \approx 1.66 \times 10^{-9}$. Since there are only 6817 genes, this is already indicative of the presence of informative genes.

2.3 Testing for Significance

With this score statistic, we can order the genes according to their potential significance. However, in choosing a subset of genes for further analysis, it is usually not clear how many genes should be included in the set. On the one hand, we would like the set to be small so that computational burden is lighter in subsequent analysis; on the other hand, we do not want to exclude information that might be contained in additional genes.

We can first see how much information is contained in the original data. We can do this by plotting the distribution of scores from the data, and comparing it to the null distribution that one would obtain if the genes were independent of the classification.

Some care must be taken when calculating the null distribution. One is tempted to compute it by generating many random sequences of n_1 0's and n_2 1's and then finding the distribution of their scores. No simple formulas

seem to exist for this distribution in general, but it is not difficult to find the number of configurations of 0's and 1's that will give a particular score. [c]

However, this is not the correct null distribution. Implicit in this approach is the assumption that all the genes are independent. This is in contrast to a well-known fact that the gene expressions are correlated among genes, a fact we also verify from the data.

To compute the null distribution while preserving the correlation structure of the data, we generate a random permutation of the entire columns, keeping all the expression levels for each patients together. From this one permutation, we calculate the distribution of the score statistics from all 7129 probes. In Figure 3, we show the distributions of the scores from the original data and from one realization of the column-permuted data. The original data gives heavier tails as expected, indicating that many genes are differentially expressed in the two classes. Next, we randomly permute the data this way many times, generating many such distributions. A p-value can then be computed by comparing the distribution from the original data to the set of distributions obtained from the randomly column-permuted data.

One way of comparing the distribution of the scores from the original data to those from the permuted data is to compute, for ith permutation, the quantity

$$S_i = \sum_{j=0}^{n_1 n_2} \left(f_i(x_j) - f_i^*(x_j)\right)^2, \quad i = 1, \ldots, M,$$

where f_i^* is the average of all distributions except the ith one,

$$f_i^*(x_j) = \frac{1}{M-1} \sum_{k=1, k\neq i}^{M} f_k(x_j),$$

and M is the number of permuted data sets ($M = 10000$ in the following). S_i is thus the measure of how much the ith distribution is different from the average of all the other distributions. We use a sum of squared differences here, but some other metric also can be considered. Because genes are not independent, it is difficult to use large sample theory common in statistics.

Based on the S_i values, the observed correlation between the expression levels and the group distinction is significant with the p-value of 0.0053 for

[c] There is yet another way of formulating the algorithm, which is helpful for computing this distribution. We consider distributing n_1 balls into $n_2 + 1$ bins. Then the score is the smallest number of moves we need in order to have all the balls on the left, with a score of 1 for each move of a ball to the next bin. We can see from this formulation that for the score k, the corresponding frequency in the distribution is the number of integer solutions $\{x_i\}_{i=1,\ldots,n_2+1}$ that satisfy $\sum_{i=1}^{n_2} i\, x_{i+1} = k$ under the constraint $\sum_{i=1}^{n_2+1} x_i = n_1$.

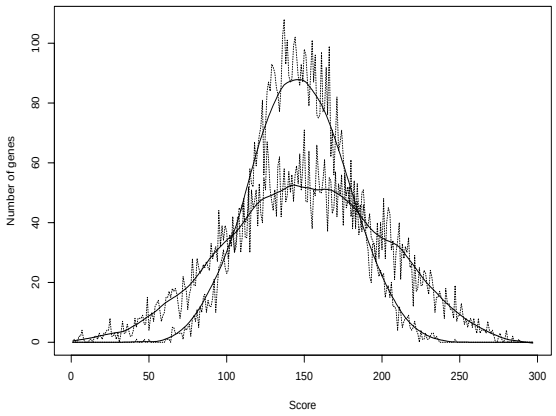

Figure 3: The distribution of scores from the original data is more spread out, with heavy tails indicative of predictive genes; the other distribution with smaller variance is one realization of randomly permuted data. The dotted lines indicate histograms and the smooth lines are obtained using robust locally linear fits.

the original data. This verifies our expectation that the data contains useful information in distinguishing the two groups. We note that this value is very large compared to 1.66×10^{-9}, which was obtained above for a single gene. Assessing the p-value for a pattern in the data must be done with caution in order to avoid gross underestimates.

In Figure 4, we show the changing p-values as the more significant genes are successively deleted from the data. When 440 genes are eliminated, the p-value goes up to 0.01; 440 is about 6% of the data. So, in the absence of other information regarding the data, 5% of the data seems to be a reasonable number for further analysis. It is true that much fewer genes may be sufficient in categorizing the samples with known classes; in fact, we saw that a single gene gives a perfect split in this data set. However, by including 5% of the genes, we are building in redundant information for more accurate and robust prediction of new samples. We find that included among the top 5% of the genes are all 50 that were chosen by Golub *et al.* (1999) using another metric for identifying informative genes. When the number of excluded genes goes beyond several hundred, we must be careful in comparing the distributions, as significance can result from the tails of the score distribution being too thin. We remedy this situation by leaving out the tails in computing S_i and rescaling

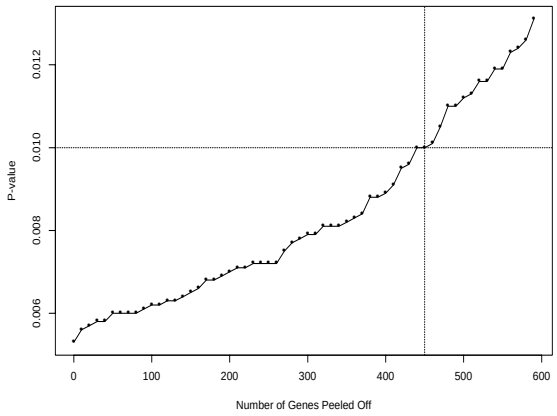

Figure 4: The p-value for the data rises as an increasingly larger number of the most significant genes are deleted. The p-values are in comparison to the 10000 randomly column-permuted samples.

the distribution appropriately before comparing.

3 Discussion

We emphasize that in order to get the p-value, we compare the distribution of the score statistics from the given data to the distributions generated by permuting entire columns while keeping the gene to gene correlation fixed. If we had assumed independence among genes, the distributions from the permuted data would look quite different, and would give us lower p-values. In Figure 5, we plotted two distributions, one assuming independence among genes (on the left), and the other without such assumption (on the right). We see that the distributions are more uniform when independence is assumed. Underestimating the p-value assigns more significance than warranted by the data and should be avoided.

The nonparametric method introduced here has several advantages. An important feature is its robustness. Because it uses ranks rather than actual expression levels, it is more robust to outliers. There is a long sequence of steps in the laboratory as well as in the image analysis before a single number is produced for an expression level, and there are many potential sources of error. That the expression levels are often unreliable is evident in the fact

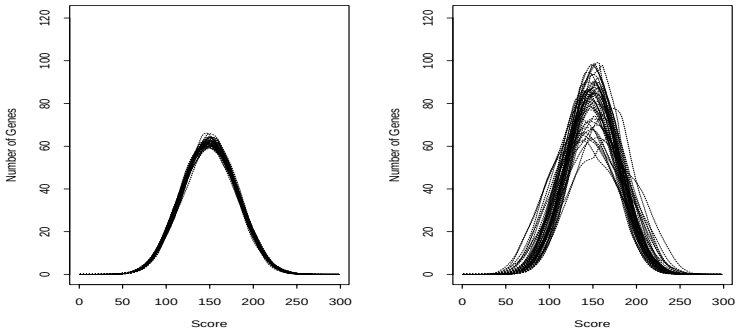

Figure 5: If the genes are assumed to be independent, the distributions are more uniform as shown on the left and the resulting p-values would be underestimated. By keeping the correlation structure for the genes, the resulting distributions are more varied. (50 distributions are plotted for each case)

that individual observations are often thrown away on the basis of poor image quality.

Microarray data are usually normalized before any analysis is carried out. There are a variety of such methods. A common rescaling, for example, is to make the overall intensities for each chip the same, in order to reduce the chip-to-chip variations.[7] However, our nonparametric method is not affected by any of the normalization steps, since order of the expression levels are left unchanged.

This fast and simple method is a formal approach for reduction of dimensionality in the data. It can be used as a filter that is more refined and systematic than those that simply look at the relative or absolute changes in fold variation. Therefore, when the classification is known for a set of samples, this nonparametric approach may be beneficial as a first step.

Related work in progress includes extensions of nonparametric approaches to other aspects of microarray data analysis, as well as correct estimates of p-values in different types of data sets.

Acknowledgments

We thank the anonymous referees for their comments. This work is supported in part by grants ST32-AI07358 for P. J. P. and R01-AI28076 for M. P. and M. B. from the National Institutes of Health.

References

1. Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J. G., Sabet H., Tran T., Yu X., *et al.*, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503-511 (2000)
2. Alon U., Barkai N., Notterman D. A., Gish K., Ybarra S., Mack D. & Levine A. J., 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 (1999)
3. Botstein D. & Brown P., Exploring the new world of the genome with DNA microarrays, *Nature Genetics (Supp.)* **21**, 33-37 (1999)
4. Butte A. & Kohane I., Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 18-29 (2000)
5. DeRisi J., Penland L., Brown P. O., Bittner M. L., Meltzer P. S., Ray M., Chen Y. D., Su Y. A. & Trent J. M., Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* **14**, 457-460 (1996)
6. Eisen M. B., Spellman P. T., Brown P. O. & Botstein D., Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868 (1998).
7. 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. & Lander E. S., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537 (1999).
8. Iyer V. R., Eisen M. B., Ross D. T., Schuler G., Moore T., Lee J. C. F., Trent J. M., Staudt L. M., Hudson J., Boguski M. S., *et al.*, The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83-87 (1999)
9. Hettmansperger T. P., *Statistical inference based on ranks.* Wiley, New York (1984)
10. Kohonen T., *Self-organizing maps.* Springer, Berlin (1997)
11. Little R. J. A. & Rubin, D. B., *Statistical analysis with missing data.* Wiley, New York (1987)
12. Ross D. T., Scherf U., Eisen M. B., Perou C. M., Rees C., Spellman P., Iyer V., Jeffrey S. S., Van de Rijn M. & Waltham M., Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**, 227-235 (2000)

13. Scherf U., Ross D. T., Waltham M., Smith L. H., Lee J. K., Tanabe L., Kohn K. W., Reinhold W. C., Myers T. G., Andrews D. T., *et al.*, A gene expression database for the molecular pharmacology of cancer. *Nature Genetics* **24**, 236-244 (2000)
14. Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E. S. & Golub T. R., Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907-2912 (1999)