

Rank-Based Analysis for the Time-course Microarray Data

Sung-Gon Yi
Department of Statistics,
Seoul National University
San 56-1. Sillim-dong,
Gwanak-gu
Seoul, Korea
skonmeme@gmail.com

Yoon-Jeong Joo
Program in Bioinformatics,
Seoul National University
San 56-1. Sillim-dong,
Gwanak-gu
Seoul, Korea
zoooyj@naver.com

Taesung Park
Department of Statistics,
Seoul National University
San 56-1. Sillim-dong,
Gwanak-gu
Seoul, Korea
tspark@snu.ac.kr

ABSTRACT

Microarray technology allows the monitoring of expression levels for thousands of genes simultaneously. In time-course microarray experiments in which gene expression is monitored over time, we are interested in clustering genes that show similar temporal profiles and identifying genes that show a pre-specified profile. Unfortunately, many traditional clustering methods used for analyzing the microarray data do not effectively detect temporal profiles for the time-course microarray data. We propose a rank based analysis for the time-course microarray data.

Our clustering method consists of two steps. The first step discretizes the expression data into groups and then transform them into the rank data. The second step performs the rank-based clustering analysis. Our testing procedure uses the bootstrap samples to select the genes that show similar patterns for the candidate profiles. Simulation study is performed to evaluate the performance of the proposed rank based method and compare it with other clustering methods. The rank based method is shown to be more powerful than other methods. The proposed method is illustrated with a breast cancer data introduced by Lobenhofer *et al.* (2002)[6].

1. INTRODUCTION

Microarray technology has initiated large-scale data based studies, such as functional genomics and systems biology, and it also has led to an explosion of molecular profiling studies which have increased the understanding of the biological mechanisms and the regulation of organisms [9, 8]. One important application of microarray technology is the time-course experiment, which focuses on studying the patterns of gene expression across a series of time points for the construction of genetic network. The development of clustering algorithms and the gene selection methods have been an important issue in functional genomics and genetic network.

However, the traditional clustering methods such as hier-

archical clustering method [2] and K-means algorithm [10], and the traditional gene selection methods like two sample t-test and its modifications [3, 1, 12] do not handle appropriately the temporal profiles of the time-course microarray data. Thus, a more specialized clustering method is required to handle the time-course microarray data appropriately. Several clustering methods have been proposed for time-course microarray data using expression profiles and slopes [13], and linear splines [4]. Order-restricted inference method was also proposed by Peddada *et al.* (2003) [7]. Genes are selected when they meet a bootstrap-based criterion for statistical significance and are assigned to the best fitting candidate profile. Quadratic regression fitting method was proposed for non-cyclic temporal expression profile [5].

In this paper, we propose the rank-based method for the analysis of time-course microarray data. The profiles of interest can be re-expressed in terms of ranks. For example, the genes showing monotone profiles have the same monotone rank profile. Also, genes showing convex or concave profile can be identified by corresponding rank profiles. The proposed method uses the a bootstrap significance test to select genes belong to the pre-specified profiles. Simulation study is performed to compare the rank based method with other clustering methods such as the method of Peddada *et al.* (2003) [7]. The proposed method is simple to use and more powerful in identifying genes than the other methods.

The paper is organized as follows. The proposed rank-based method and the test procedure are presented in Section 2. Simulations results are given in Section 3. In addition, the proposed method is illustrated with a breast cancer data introduced by Lobenhofer *et al.* (2002) [6]. Finally, the concluding remarks are summarized in Section 4.

2. RANK BASED METHOD

Suppose there are T time points denoted by $t(= 1, \dots, T)$ and there are G genes denoted by $g(= 1, \dots, G)$. Let x_{gt} represent the expression level for the g th gene at the t th time point. Then, all microarray experiment data for the g th gene can be represented using a vector, (x_{g1}, \dots, x_{gT}) . Let $(\mu_{g1}, \dots, \mu_{gT})$ be the corresponding mean vector. The temporal profiles can be defined in terms of means. The common profiles considered by Peddada *et al.* (2003) [7] are monotone increasing profile, monotone decreasing profile, convex (up-down) profile, and concave (down-up) profile.

Without missing observations, there are T observations for the g th gene. If there are no tied observations, each observation has its unique rank ranging from 1 to T . Let

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the VLDB copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Very Large Data Base Endowment. To copy otherwise, or to republish, to post on servers or to redistribute to lists, requires a fee and/or special permission from the publisher, ACM.

VLDB '07, September 23-28, 2007, Vienna, Austria.
Copyright 2007 VLDB Endowment, ACM 978-1-59593-649-3/07/09.

(r_{g1}, \dots, r_{gT}) be the rank vector of $(x_{g1}, x_{g2}, \dots, x_{gT})$. For each mean profile, there is a corresponding rank profile. Instead of mean values, we can use the corresponding ranks to define rank profiles. For example, the up-down profile with maximum at time i can be derived as follows:

$$C = \{r \in R^T : r_1 \leq \dots \leq r_{i-1} \leq r_i \geq r_{i+1} \geq \dots \geq r_T\}. \quad (1)$$

All other rank profiles can be defined similarly.

Our proposed method consists of two-steps. The first step is discretization and the second step is a bootstrap test for identifying genes that show a specific candidate profile.

2.1 Two-step discretization

When the original microarray data are transformed directly into the ranks, two or more ranks can be easily reversed on temporal profile. For example, suppose that there is an expression profile (0.99, 1.1, 2.1, 3.2, 4.3, 0.98). Here 0.98, 0.99, and 1.1 are very similar expression levels. However, when we transform the expression data into the ranks directly, they become (2, 3, 4, 5, 6, 1), and the values of 1st, 2nd, and 6th points can be treated as quite different values.

To avoid this problem of rank transformation, we introduce a discretization step before rank transformation. This discretization step is to classify the observed expression data into the several blocks. For example, the discretization step classify the original data (0.99, 1.1, 2.1, 3.2, 4.3, 0.98) into the four groups: (0.98, 0.99, 1.1), (2.1), (3.2), and (4.3). Then, the new rank profile becomes (2, 2, 4, 5, 6, 2) using the mid-rank. The expression values within the same block are treated as having tied ranks.

We propose the two types of discretization methods: percentile discretization and equal interval discretization. The percentile discretization method constructs the blocks using the percentiles. For a given percentile value q , the total range of expression is divided into the several blocks, where each block contains $100q\%$ of the data. The equal interval discretization method constructs the blocks of equal size. First compute the standard deviation (SD) of the whole expression values. Let m be a parameter for defining the block size. Then, the interval of equal length $2 \cdot SD/m$ is defined from the center point.

For applying the rank-based algorithm, the blocks are defined first by applying one of the discretization methods, and then the expression data are transformed into the ranks.

2.2 Bootstrap test

We propose a bootstrap test to classify the genes into the one of the candidate profiles. First, we transform the expression data into the ranks by using one of the discretization methods. Then, we calculate the distance between the gene profile and the candidate profiles, and classify the genes into the candidate profile which has the minimum distance.

The detailed steps are given below.

- Step 1 Pre-specify a collection of candidate profiles C^1, C^2, \dots, C^p in rank.
- Step 2 Compute the rank vector for the g th gene, using two step discretization.
- Step 3 Compute the Euclidian distance between the rank vector of the g th gene and the candidate profile C^i for $i = 1, \dots, p$. When there are more than one sub-profiles, the distance, $D(g, C^i)$, can be defined as the minimum of the distance between gene and sub-profiles.

Step 4 (Bootstrap null distribution and significance test) Assume that the true location and scale parameters are the same at every time point. Draw B bootstrap samples of size T by a simple random sampling with replacement from original samples after ignoring the time points, and calculate $D(g^b, C^i)$, $b = 1, \dots, B$.

To test whether the profile of the g th gene is a null profile or not, use the bootstrap distribution. Suppose that $D(g, C^r)$ is the smallest distance. Then, assign gene g to the profile to C^r if $D(g, C^r)$ is less than a threshold from the bootstrap distribution. Otherwise, do not classify g into any of the profile.

3. RESULTS

3.1 Simulation

We performed a simulation study to evaluate the performance of rank based methods. This profile has three time points and four replications for each time point. There are 940 null profiles and 60 temporal profiles. Temporal profiles consist of profiles such as monotone increasing, monotone decreasing, up-down with minimum at time 1, and down-up with minimum at time 1. We set the mean vector for the monotone increasing profile as (-1, 0, 1), for the monotone decreasing profile as (1, 0, -1), for the up-down profile as $\{(-1, 1, 0), (0, 1, -1)\}$, and for the down-up profile as $\{(1, -1, 0), (0, -1, 1)\}$. The null profile is set as (0, 0, 0). We generated the eight dataset from the normal distributions with SDs 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, and 1, respectively.

In Table 1, we evaluated performances of Peddada *et al.* (2003) [7]'s method, the original rank based method denoted by Rank1, the rank profile after percentile discretization, denoted by Rank2, and the rank profile after equal interval discretization method, denoted by Rank3. Four methods were compared using sizes, powers, and the correct profile detection rates. The size, defined as the proportion of the false positive error, were computed as the proportions of the selected profiles out of 940 null profiles. The powers were computed in two ways. The first one was computed as the proportions of selecting any profile from the 60 candidate temporal profiles. The second one was computed as the proportion that the selected profile C^r is the true temporal profile. In Table 1, the first one is denoted as power and the second one is denoted by the correct profile detection rate. We used the significance level, α , of 0.025. The simulation results are summarized in Table 1. We obtained the results for the various values of q for percentile discretization and m for equal interval discretization. Here, the results were given for the cases where the value of q was 0.05 for percentile discretization and the value of m was 3 for equal interval discretization.

The sizes of Rank1 and Rank2 were all greater than 0.025 showing the inflated type I errors. Therefore, many false positive genes might be detected. These high false positive errors may be due to the high variabilities of the expression values, causing the small differences to lead to the different ranks. In particular, the performance of Rank2 was not better than the original rank based method, Rank1. That is, the percentile discretization seemed not effective.

The sizes of Peddada were small, but its statistical powers decreased rapidly, as the SD increased. It is probably due to

Table 1: The results of simulation Case 1.

	SD	Peddada ^a	Rank1 ^b	Rank2 ^c	Rank3 ^d
size	0.125	0.0053	0.1564	0.1468	0
	0.25	0.0106	0.1543	0.1553	0
	0.375	0.0096	0.1691	0.15	0.0011
	0.5	0.0128	0.1511	0.1521	0.0096
	0.625	0.0128	0.1340	0.1638	0.0096
	0.75	0.0117	0.15	0.1617	0.0085
	0.875	0.0117	0.1309	0.1734	0.0160
	1	0.0106	0.1468	0.1670	0.0085
power	0.125	1	0.1667	0.1468	1
	0.25	1	0.1667	0.1553	1
	0.375	1	0.1833	0.15	0.9833
	0.5	0.8667	0.2	0.1521	0.8667
	0.625	0.7333	0.15	0.1638	0.8167
	0.75	0.4667	0.1333	0.1617	0.5833
	0.875	0.3333	0.1	0.1734	0.5
	1	0.2333	0.16667	0.1670	0.4
correct profile detection rate	0.125	1	1	1	1
	0.25	1	1	1	1
	0.375	1	1	1	1
	0.5	1	1	1	1
	0.625	0.9318	0.8889	1	1
	0.75	0.8571	1	1	1
	0.875	0.85	0.8333	0.9286	0.9333
	1	0.7857	0.9	0.9375	0.9167

^a Peddada *et al.* (2003)'s method

^b Original rank based method

^c Rank based method after the percentile dependent discretization. q is 5%.

^d Rank based method after the equal interval discretization. m is 3.

the fact that the deviance of the statistic used in Peddada becomes extremely large when the data showed a large SD.

Unlike other two rank based methods, Rank3 preserved its sizes less than 0.025 and its power was equal to or greater than that of Peddada's method. When SD is small, Peddada and Rank3 showed similar powers. On the other hand, as SD increased, power of Rank3 became larger than that of Peddada. As SD increased, this differences of powers between two methods increased. The correct profile detection rates showed similar patterns to those of powers.

3.2 Breast cancer cell line data of Lobenhofer *et al.* (2002)

Lobenhofer *et al.* (2002) [6] had performed the study of MCF-7 breast cancer cell line treated with 17 β -estradiol or ethanol (vehicle control). Samples were cultured at 1, 4, 12, 24, 36, and 48 hours after 17 β -estradiol or ethanol was treated. At each time point sample was biologically replicated and each sample was hybridized to four microarrays. Therefore, there were total 8 microarrays at each time point. NIEHS Human ToxChip v1.0, was used for microarray experiment. There were 1936 spots on each chip.

Ten candidate profiles were considered, as shown in Table 2. They include one monotone increasing profile, one monotone decreasing profile, four convex profiles, and four concave profiles. We analyzed data by Peddada and Rank3. Rank1 and Rank2 were excluded from the analysis because of their poor performance in simulation studies. 121 genes were selected by Peddada and 130 genes were selected by

Table 2: 10 types of candidate profile and number of selected genes on each profile

Profile	Id	Number of genes (Peddada)(Rank3)	
Monotone	Increasing	1	16
	Decreasing	2	5
Concave	Maximum at time 2	3	1
	Maximum at time 3	4	12
	Maximum at time 4	5	7
	Maximum at time 5	6	56
Convex	Minimum at time 2	7	6
	Minimum at time 3	8	10
	Minimum at time 4	9	7
	Minimum at time 5	10	1
			13

Rank3 at a significance level 0.0025. We used the multiplier $m = 1.3$ for Rank3.

Table 2 shows the number of genes clustered to each candidate profile. Figure 1 shows the mean expression levels for the candidate profiles. Selected genes by two methods were quite different. The number of common genes selected by both methods are only 44.

Figure 1 shows the mean profiles. The profiles of genes selected by Peddada tended to show larger fluctuations than Rank3, as shown in Figure 1a. For example, although both methods selected the same number of genes for Profile 2, Rank3 yielded much smaller fluctuations than Peddada.

It seemed that Rank3 removed most genes with large fluctuations and kept genes with small fluctuations. Figure 2 shows the rank profile of selected genes by Rank3 which show similar profiles to those of candidate profiles.

We additionally clustered one more specific profile, decreasing after 4th time point, $\{\mu \in R^6 : \mu_1 = \dots = \mu_4 \geq \mu_5 \geq \mu_6\}$. This profile can be represented by (4.5, 4.5, 4.5, 2, 1) as rank profile. Specific profiles with tied levels like that are hardly to be analyzed by Peddada. 31 genes were selected by (Rank3) when α was 0.0025 and m was 1.3. Figure 3 shows profile plots of selected genes. It seems profiles of selected genes were represents well.

4. DISCUSSION

In this paper, we propose a rank based method for the time-course microarray data. It first discretized data for clustering and gene selection. The proposed method uses the ranks defined over all time points. To improve the performance, additional discretization is introduced before rank transformation. For the clustering analysis, the genes are clustered using the discretized ranks. Temporal profiles of interest are defined in terms of inequalities among mean levels. The corresponding rank profiles can then be easily derived from the mean temporal profiles.

The proposed method is similar in spirit to that of Peddada *et al.* (2003) [7]. Instead of using time-consuming order-restricted inference, we applied a rank-based method. We compared the performance of the rank based method with that of Peddada *et al.* (2003) [7]. We showed that both methods preserved significance levels well but our proposed method yielded larger power. Especially, the performance of the rank based method was much better when the microarray data contained large variabilities.

We analyzed the breast cancer cell line data of Lobenhofer

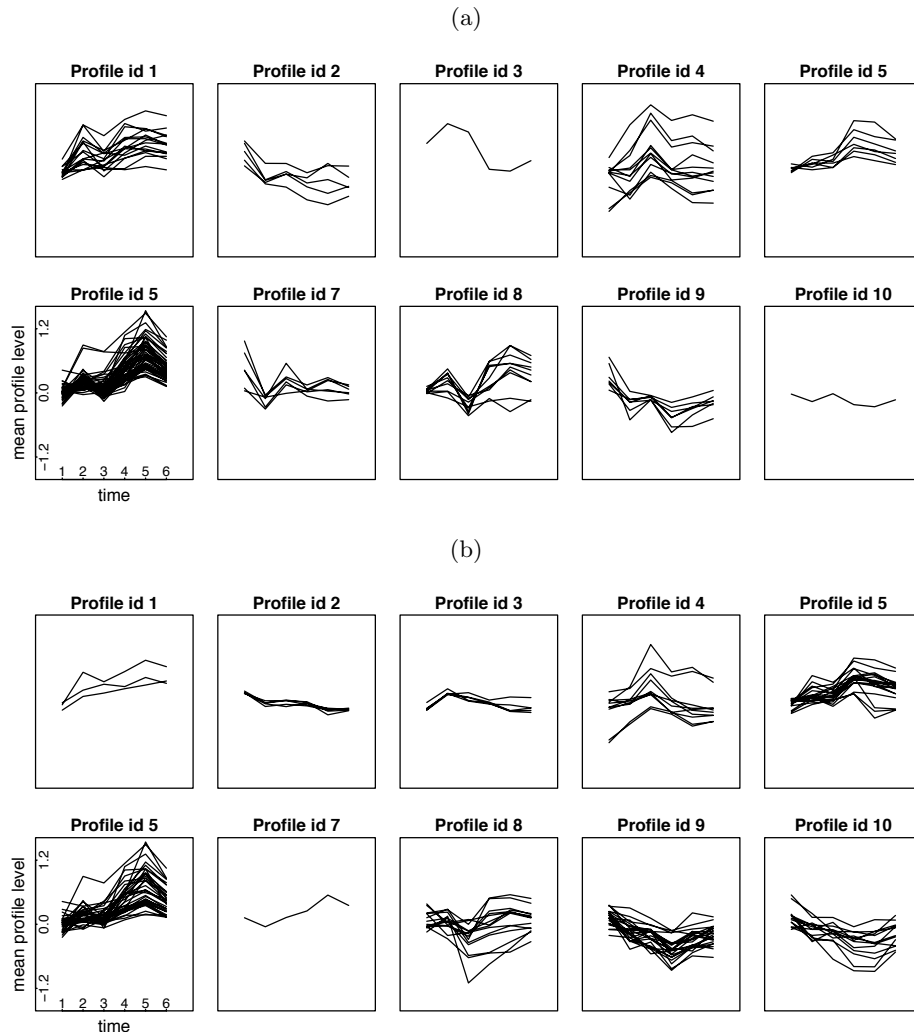


Figure 1: Mean expression profiles. (a) 121 selected genes using Peddada (b) 130 selected genes using Rank3

et al. (2002) [6] and selected 130 genes by the rank based method. These genes were divided into ten different profiles reasonably well. In addition, we demonstrated that Rank3 could easily identify 31 genes for a specific candidate profile which could not be analyzed by Peddada.

Microrarray data contain many missing observations. Before applying the rank-based approach, these missing observations need to be imputed first. The most common imputation method is the K-nearest neighbors method [11]. After imputing missing observations, the rank-based method can be applied.

Acknowledgments

The work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126).

5. REFERENCES

- [1] P. Baldi and A. D. L. and. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.
- [2] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
- [3] T. Golub, D. Slonim, P. Tamayo, C. Huard, J. Caasenbek, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
- [4] D. Hoon, S. Imoto, and S. Miyano. Statistical analysis of a small set of time-ordered gene expression data using linear splines. *Bioinformatics*, 18:1477–1485, 2002.
- [5] H. Liu, S. Tarima, A. S. Borders, and T. V. Getchell. Quadratic regression analysis for gene discovery and pattern experiments. *BMC Bioinformatics*, 6:106, 2005.
- [6] E. K. Lobenhofer, L. Bennett, P. L. Cable, L. Li, P. R.

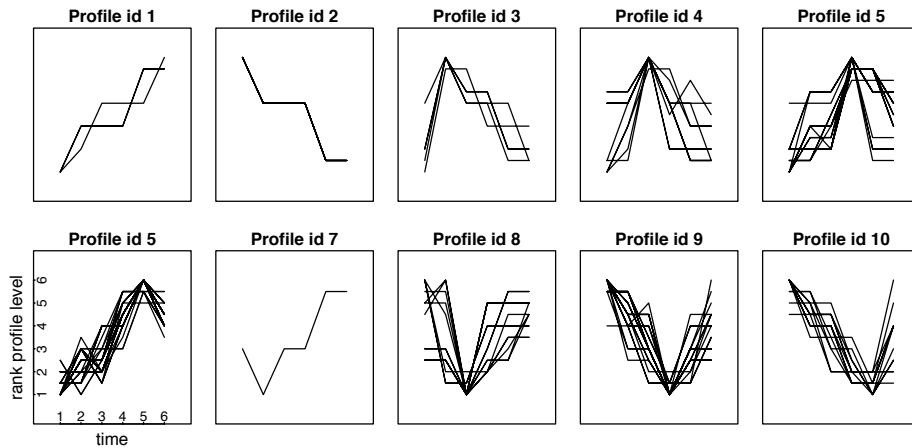


Figure 2: Rank profiles of 130 selected genes using Rank3 with $m = 1.3$

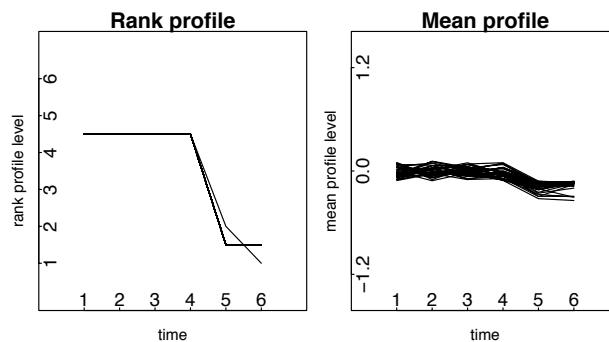


Figure 3: The profile of selected genes from a specific candidate, decreasing after 4th time point.

- Bushel, and C. A. Afshari. Regulation of dna replication fork genes by 17β -estradiol. *Molecular Endocrinology*, 16:1215–1229, 2002.
- [7] S. Peddada, E. Lobenhofer, L. Li, C. Afshari, C. Weinberg, and D. M. Umbach. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics*, 19:834–841, 2003.
- [8] M. Schena. *DNA microarrays: a practical approach*. Oxford University Press, 1999.
- [9] M. Schena, D. Shalon, R. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.
- [10] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22:281–285, 1999.
- [11] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17:520–525, 2001.
- [12] V. Tusher, R. Tibshirani, and G. Chu. significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121, 2001.
- [13] X. Wen, S. Fuhrman, D. Carr, S. Smith, J. Barker, and R. Somogy. Large scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA*, 95:334–339, 1998.