PAPER

Extraction of Informative Genes from Multiple Microarray Data Integrated by Rank-Based Approach

Dongwan HONG[†,††], ***Nonmember***, **Jeehee YOON**[†a)], ***Member***, **Jongkeun LEE**[†], **Sanghyun PARK**[†††], ***and*** **Jongil KIM**[††], ***Nonmembers***

SUMMARY By converting the expression values of each sample into the corresponding rank values, the rank-based approach enables the direct integration of multiple microarray data produced by different laboratories and/or different techniques. In this study, we verify through statistical and experimental methods that informative genes can be extracted from multiple microarray data integrated by the rank-based approach (briefly, integrated rank-based microarray data). First, after showing that a non-parametric technique can be used effectively as a scoring metric for rank-based microarray data, we prove that the scoring results from integrated rank-based microarray data are statistically significant. Next, through experimental comparisons, we show that the informative genes from integrated rank-based microarray data are statistically more significant than those of single-microarray data. In addition, by comparing the lists of informative genes extracted from experimental data, we show that the rank-based data integration method extracts more significant genes than the z-score-based normalization technique or the rank products technique. Public cancer microarray data were used for our experiments and the marker genes list from the CGAP database was used to compare the extracted genes. The GO database and the GSEA method were also used to analyze the functionalities of the extracted genes.

key words: *microarray data, data integration, informative gene, significance test*

1. Introduction

It is possible to obtain a large amount of gene expression data by performing a single microarray experiment. Therefore, microarray experiments are useful for identifying the phenotypes of diseases. In general, as microarray experiments have more samples, we can expect more reliable and valid results. However, microarray experiments are expensive. Therefore, it is not easy to perform microarray experiments with a large number of samples. In practice, microarray experiments with the same goal may produce different results if the people performing the experiments or the experimental environments vary. In addition, errors are likely to arise in microarray experiments.

Recently, microarray experimental data have accumulated rapidly. Therefore, research involving the integration of existing microarray data to increase the sample size and, accordingly, to increase the reliability is becoming more popular. In the earlier studies [1], [2], we used a rank-based approach to develop techniques for integrating microarray data and building phenotype classifiers. The rank-based approach converts the expression values of each sample into the corresponding rank values within the sample. It enables the integration of multiple sets of microarray data produced in different laboratories or by different techniques.

In the present paper, we verify through statistical and experimental methods that informative genes can be extracted from multiple microarray data integrated by the rank-based approach. To attain this goal, we first show that the non-parametric technique can be used effectively as a scoring metric for rank-based microarray data and that the scoring results from integrated rank-based microarray data are statistically significant. Next, through experimental comparisons, we show that the informative genes from integrated rank-based microarray data are statistically more significant than those from single-microarray data. Lastly, by comparing the rank-based data integration method with the existing data integration methods, we verify that the informative genes from integrated rank-based microarray data are more significant than those from microarray data as integrated by the existing methods.

We employ *Park*'s non-parametric method [3] as a scoring metric for the rank-based microarray data. This method calculates the score of each gene by comparing its expression values across all samples. Therefore, the scoring results from microarray data of actual gene expression values may be different from the scoring results from its corresponding rank-based microarray data. We performed permutation tests to verify the statistical significance of the scoring results from the rank-based microarray data. In the tests, the scoring results from the integrated rank-based microarray data were found to be statistically significant.

Next, using an experimental approach, we prove that informative genes can be extracted from integrated rank-based microarray data. We used public prostate cancer microarray data for the experiments along with the marker genes list from the CGAP (Cancer Genome Anatomy Project, http://cgap.nci.nih.gov) database [4] for the assessment and comparison of extracted genes. In more detail, we assessed and compared how many genes in the list of informative genes extracted from each experiment data are also contained in the marker genes list from the CGAP database. In addition, we analyzed the function-

Manuscript received March 30, 2010.
Manuscript revised October 15, 2010.

[†]The authors are with the Department of Computer Engineering, Hallym University, Korea.
[††]The authors are with the Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Korea.
[†††]The author is with the Department of Computer Science, Yonsei University, Korea.
a) E-mail: jhyoon@hallym.ac.kr

DOI: 10.1587/transinf.E94.D.841

alities of the informative genes extracted from each experiment data. To understand the functional associations between the genes and cancer-related activities, we first retrieved the associated functions of the genes by looking them up in the GO (Gene Ontology) database [5] and then classified and evaluated them biologically. Moreover, by using the GSEA (Gene Set Enrichment Analysis, http://www.broadinstitute.org/gsea) method [6], [7] that evaluates microarray data at the level of gene sets, we tested enrichment of cancer-related gene sets included in the subcatalog CM (Cancer Modules) of the MSigDB database, and evaluated the clinical annotations of the top scoring gene sets.

The remainder of the paper is organized as follows. In Sect. 2, we briefly describe existing approaches for the integration of microarray data and the extraction of informative genes. In Sect. 3, we explain rank-based approaches for the integration of microarray data, the gene scoring method, and the extraction of informative genes. In Sect. 4, through extensive experiments with publicly available microarray data, we show the usefulness of the rank-based integration of microarray data and verify that informative genes can be extracted from integrated rank-based microarray data. Finally, in Sect. 5, we conclude this paper with the presentation of a couple of future research directions.

2. Related Work

Experimental microarray data are organized as matrices where the rows are the genes and the columns are the samples. The integration of multiple microarray data is not easy in most cases; although they have the same research goals, their platforms, protocols, gene sets, and scales of expression values may be different. Typical methods used for the integration of multiple microarray data include the meta-analysis method [8], [9], the normalization and transformation method [10], [11], and the rank-based approach [1], [12].

Instead of comparing actual expression values of individual microarray data, the meta-analysis method combines the results of individual microarray experiments using a statistical technique [8]. The rank products technique was recently proposed as a type of the meta-analysis method [9]. This technique first determines the gene orders within each sample by utilizing the fold ratio values of each experiment, and then determines the overall gene orders by multiplying the corresponding individual gene orders. However, the number of samples in a single microarray experiment is small in many cases, and the results from a single microarray experiment with a small number of samples are therefore likely to be unreliable. Therefore, the combination of such unreliable results may bring an even worse result.

The normalization and transformation method integrates a set of microarray data by converting the gene expression values of each microarray data into the corresponding values of a common scale [10]. A classical method is the z-score transformation [11], which normalizes the expression values of each sample using its mean and standard deviation values. References [10], [13] proposed the QD (Quantile Normalization) methods that transform gene expression values of different microarray data to the values within a common numerical range. The NLT (Normalized Linear Transform) method proposed recently as one of the QD methods preserves the relative ranking order of the expression values for each gene without information loss [14]. To detect significant changes in gene expressions, statistical tests such as the fold ratio, z-ratio/test [11], and t-statistical test [15] can be applied directly to normalized microarray data. However, there is still no consensus on the best way to normalize expression values.

The rank-based approach converts the expression values of each sample into the corresponding rank values. Previous work has shown that considering a gene's rank within a sample instead of its expression value eliminates systematic biases better and thus improves the classification accuracy [2], [12], [13], [16]. In reference [16], Liu et al. studied whether a direct replacement of expression values by the corresponding rank values is effective or not in cross-generation and cross-laboratory comparisons. However, they did not use the rank-based approach to integrate the microarray data. Xu et al. proposed a top-scoring pair classifier to select marker genes from integrated rank-based data [12]. However, their classifier is built simply by comparing the relative smallness and largeness of the expression values within each sample. In our previous work [1], [2], we proposed a two-stage approach for phenotype classification that integrates microarray data by rank-based approach and builds a classifier, and showed that our system has better classification accuracy, sensitivity, and specificity in the classification of an independent test dataset. However, to evaluate the microarray integration methods, their effect on gene ranking and informative gene selection has to be further investigated statistically and biologically.

One of the difficulties in the analysis of microarray data is the high dimensionality incurred by a large number of genes. However, only a small fraction of genes are informative for predicting significant changes in gene expressions. Various methods have been proposed to select informative genes precisely and effectively.

Typically, informative genes are selected according to test statistics. To represent the data being analyzed, such parametric methods as t-statistics [15], *Fisher*'s method [17], and *Golub*'s method [18] assume a statistical model. There are also non-parametric methods such as TNom [19], Wilcox rank sum [20], and *Park*'s method [3]. The non-parametric methods establish a minimum boundary and calculate the distance from the boundary as a score. On the other hand, when genes are considered as features, the rank-based feature selection method [21] can be used. This method measures the significance of features and then ranks them according to their significance. The significance of features is measured by the Information Gain [21], Relief-F [22], Hypernetworks model-based method [23], or by a method using *Kendall*'s Correlation Coefficient [24]. All of

these methods, however, use the actual expression values of each gene without considering the integration and normalization of the microarray data.

3. Methods

In this section, we briefly explain how to integrate multiple microarray data using the rank-based approach and describe the methods for gene scoring and informative gene extraction.

3.1 Rank-Based Microarray Data Integration

The rank-based method for integrating multiple microarray data is summarized as follows. First, from a set of microarray data that were generated independently but had the same experimental objectives, only the genes common to all microarray data are extracted. The expression values of each sample are then converted to the corresponding rank values within the sample. Once all expression values are changed to the corresponding rank values, the integration of samples from different experiments becomes feasible, as long as their gene order is identical. This method is simple and useful for integrating a large number of microarray samples while not requiring any normalization. Hereafter, for simplicity, experimental data with original expression values are termed *raw data*, and experimental data with rank values are termed *rank data*.

Figure 1 shows our data integration method using an example. Let us consider two microarray data sets, Data (A) and Data (B), which consist of three normal and three cancer samples, respectively. As shown in Fig. 1, the scale of the expression values of Data (A) is quite different from that of Data (B); thus, these two sets of data cannot be integrated directly. We first extract five common genes and convert all expression values into the corresponding rank values within each sample, thus giving Data (A′) and Data (B′) of the *rank data*. The two data sets are then directly merged as Integrated_Data, which consists of six normal and six cancer samples.

As integrated data contain only rank values rather than actual expression values, there may be a slight loss of information. However, expression values that are too big or too small can be noise, which may have a negative effect on the extraction of informative genes. In return, we gain robustness to external factors such as noise.

3.2 Gene Scoring and Informative Gene Extraction

Park's non-parametric scoring method [3] is extended and applied to integrated microarray data. *Park*'s method, which was originally proposed for single-microarray data, builds a binary sequence for a gene and calculates its score by measuring how differently the gene is expressed in two class

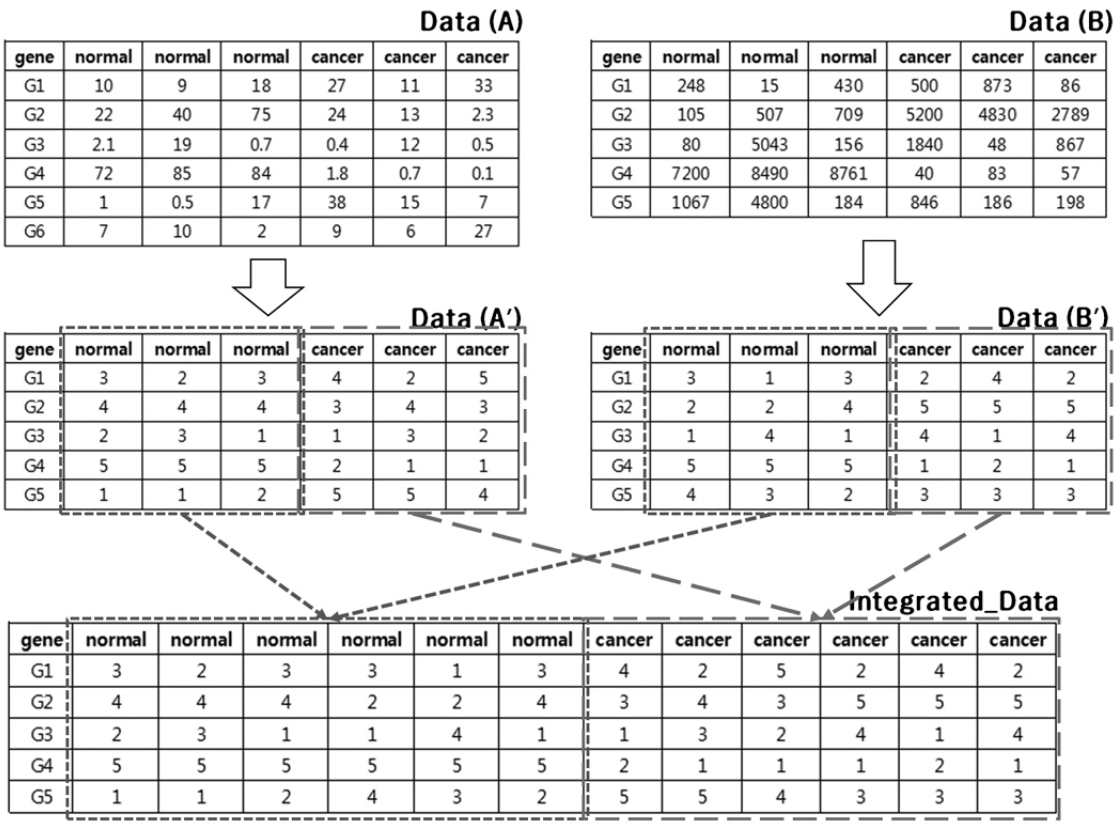


Fig. 1 An example of microarray data integration.

groups (normal and cancer groups), using *Kendall*'s Correlation Coefficient [24]. When there are n_1 normal samples and n_2 cancer samples, the score ranges from 0 to $n_1 \times n_2$. Both low and high scores indicate differentially expressed genes, which are finally selected as informative genes.

Here, we describe the informative gene extraction algorithm using two sets of data with different origins as the example. For simplicity, we assume that the two data sets contain the same gene set. Let N and C denote two sample classes (normal and cancer classes). There are A_n normal and A_c cancer samples in the first data and B_n normal and B_c cancer samples in the second data.

Algorithm 1:

Input : two microarray sets of data, the number of common genes M, the number of informative genes iM

Output : iM informative genes

1 For each sample, convert the expression values into the rank values within the sample. Merge the samples of the two sets of data, and obtain the integrated rank-based data ID[i][j] (i=1, ..., M, j=1, ..., $A_n + B_n + A_c + B_c$) where the former $A_n + B_n$ columns are for normal samples and the latter $A_c + B_c$ columns are for cancer samples.

2 Generate a sequence S that represents the class label of each sample. Here, class label 0 is assigned to a normal sample, and class label 1 is assigned to a cancer sample.

3 Determine the score Sc[i] (i=1, ..., M) for each gene using the following steps.
3.1 For all j, sort ID[i][j] in ascending order, and generate a binary sequence T where normal samples are replaced with 0's and cancer samples are replaced with 1's.
3.2 As a score, calculate the minimum number of swaps of neighboring 0 and 1 needed for transforming T into S.

4 Select the lowest iM/2 scores and the highest iM/2 scores. Return the genes with the selected scores as the informative genes.

Note that, by executing the merge step of the above algorithm iteratively, we can extract informative genes from more than two microarray data sets. If a gene has the same rank for the samples with different class labels (i.e., "normal" or "tumor"), it must not be included in a set of informative genes. Therefore, we ignore such gene by skipping the calculation of its score in step 3.2.

Let us explain the scoring method using an example. Figure 2 illustrates how the score of a gene is calculated with six sample data, 3, 1, 3, 2, 4, and 2. Here, we assume that each sample data represents a rank value. In this figure, samples 1, 2, and 3 are from a normal class and samples 4, 5, and 6 are from a cancer class. First, class label 0 is assigned to the normal samples and class label 1 is assigned to the cancer samples. Thus, we obtain an initial binary sequence S = 000111, which represents the class labels of the six sample data. Next, the sample data are sorted in ascending order along with their class labels. As a result, the sorted binary sequence T = 011001 is obtained, which now represents the class labels of the sorted sample data. The distance between T and S is defined as the minimum number

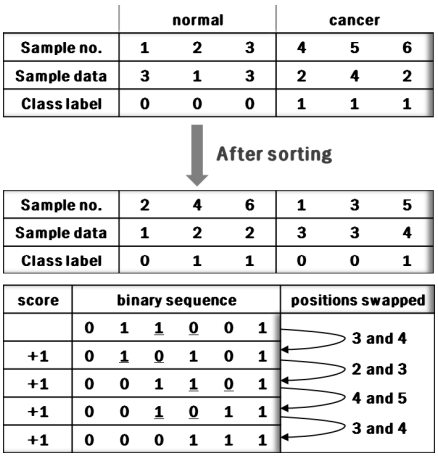


Fig. 2 An example of gene scoring.

of swaps of neighboring 0 and 1 needed to transform T into S. Figure 2 shows the process of transforming T = 011001 into S = 000111. As four swaps are required to transform T into S, their distance is defined as 4.

Next, we describe our informative gene selection procedures using the previous example given in Fig. 1. As explained previously, the score refers to the minimum number of swaps necessary to arrive at perfect splitting, with all 0's on the left and all 1's on the right. Figure 3 shows the scoring results of each data. If only two genes are selected as informative genes from each data, the gene with the highest score and the gene with the lowest score are selected. For instance, G1 and G4 are selected from Data (A), and G2 and G4 are selected from Data (B). In the same manner, G4 and G5 are selected from Data (A′), and G2 and G4 are selected from Data (B′). Note that the informative genes from the *raw data* are not identical to those from the *rank data*. Next, Data (A′) and Data (B′) are integrated and finally G4 and G5 are selected from the Integrated_Data as informative genes.

4. Experiment and Verification

This section verifies that informative genes can be effectively and efficiently extracted from integrated rank-based microarray data. In Sect. 4.1, we describe the data sets used in our experiments. In Sect. 4.2, we show the statistical significance of the scoring results of integrated rank-based data. In Sect. 4.3 we prove the superiority of the rank-based data integration method through the experimental comparisons.

4.1 Data Set

For our experiments, we chose the following two microarray data sets of cancer research.

Leukemia data set: We used the leukemia data analyzed by Golub et al. [18]. *Golub*'s data contains 38 bone marrow samples obtained from acute leukemia patients. 27 samples are from the ALL (Acute Lymphoblastic Leukemia)

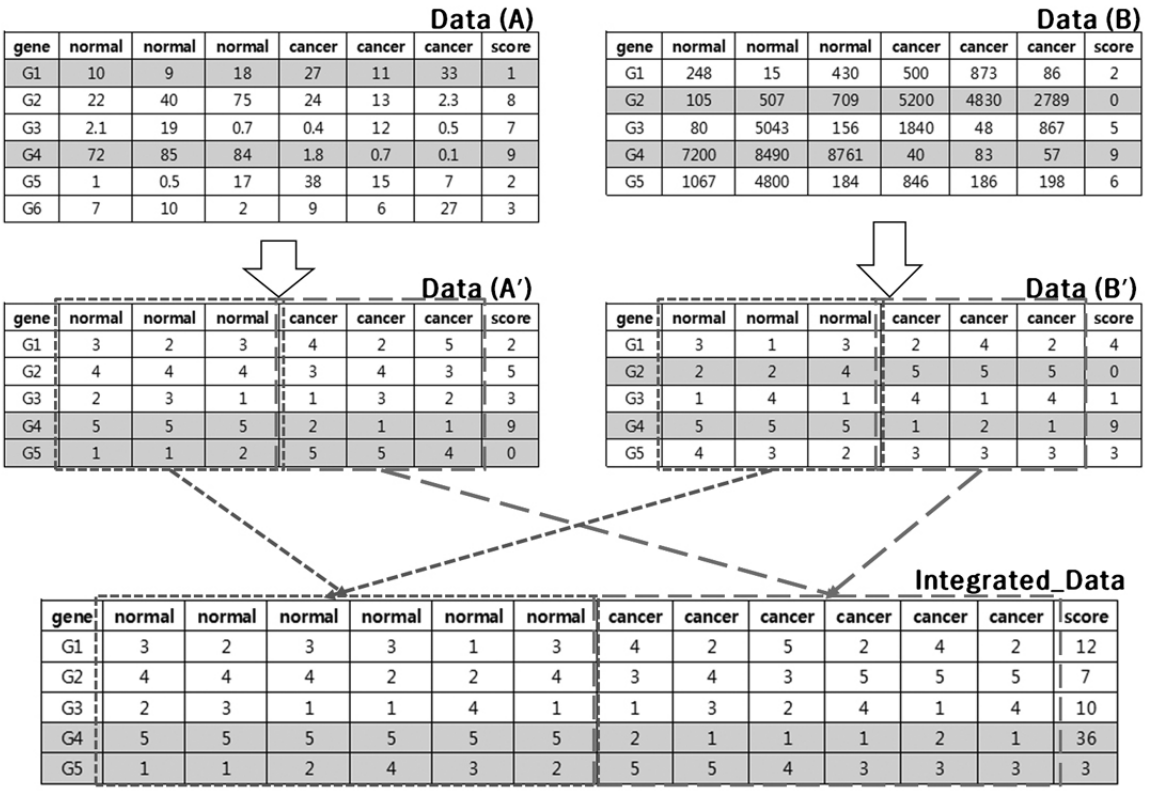


Fig. 3 An example of microarray data integration and informative gene selection.

class and 11 samples are from the AML (Acute Myeloid Leukemia) class. High-density oligonucleotide microarrays (produced by Affymetrix) containing 7,129 probes for 6,817 human genes were used.

Prostate Cancer data set: We used three prostate cancer microarray data that are publicly available. The platform of these data is Affymetrix HG 95Av2, which contains 20,180 probes for 12,600 human genes. Each data is represented as the abbreviation of the first author of the paper, as in LaTulippe [25], Welsh [26], and Singh [27]. LaTulippe consists of three normal samples, 14 primary prostate cancer samples, and nine metastatic prostate cancer samples. Welsh consists of nine normal samples and 25 cancer samples, and Singh consists of 50 normal samples and 52 cancer samples.

4.2 Significance Test

A permutation test was performed to test the significance of the gene scoring result for the *rank data*. We generated a random permutation of entire columns, keeping all the rank values for each sample together. A p-value was then computed by comparing the distribution obtained from the original data to the set of distributions obtained from the randomly permuted data.

To calculate a p-value, we used a cumulative function S_i of (Eq. (1)), where a sum of squared differences is calculated. For the comparison, we used the same function that is given in [3]. Let n_1 and n_2 denote the numbers of normal and cancer samples of the data, respectively. S_i is the measure of how much the i-th score distribution differs from the average of all the other score distributions. Here, f_i represents the score distribution of the i-th column-permuted data, and f_i^* represents the average of all distributions except for the score distribution of the i-th column-permuted data. S_0 represents the difference between the score distribution of the original data and the average of the score distributions of the other column-permuted data. A significance probability $p(S_i \geq S_0)$ is now calculated. Here, the requirement of $i = 1, \ldots, M$ is met, with M representing the number of column-permuted data. If the p-value is smaller than the significance level, we assumed that the gene scoring result is significant.

$$S_i = \sum_{j=0}^{n_1 n_2} (f_i(x_j) - f_i^*(x_j))^2, i = 1, \ldots, M$$
$$f_i^*(x_j) = \frac{1}{M-1} \sum_{k=1, k\neq i}^{M} f_k(X_j) \quad (1)$$

We applied the non-parametric scoring method described in Sect. 3.2 to *Golub*'s leukemia data [18]. A permutation test was performed for the two sets of data, the *raw data* and the *rank data*. We performed 10,000 permutations. Figure 4 shows the score distributions from the original data and from a set of randomly column-permuted data. The score is given along the x-axis, ranging from 0 to 297, and the number of genes having each score is plot-

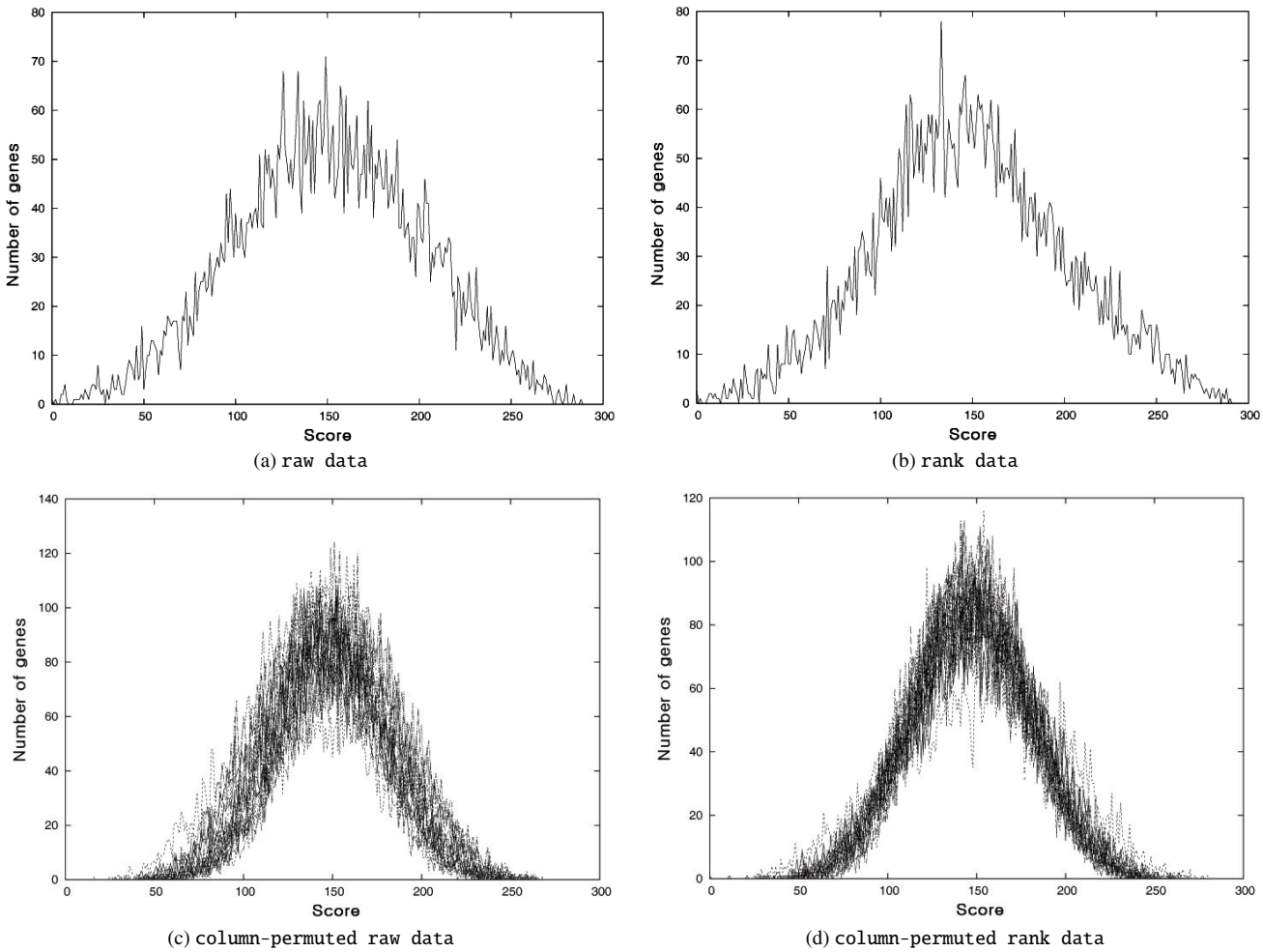
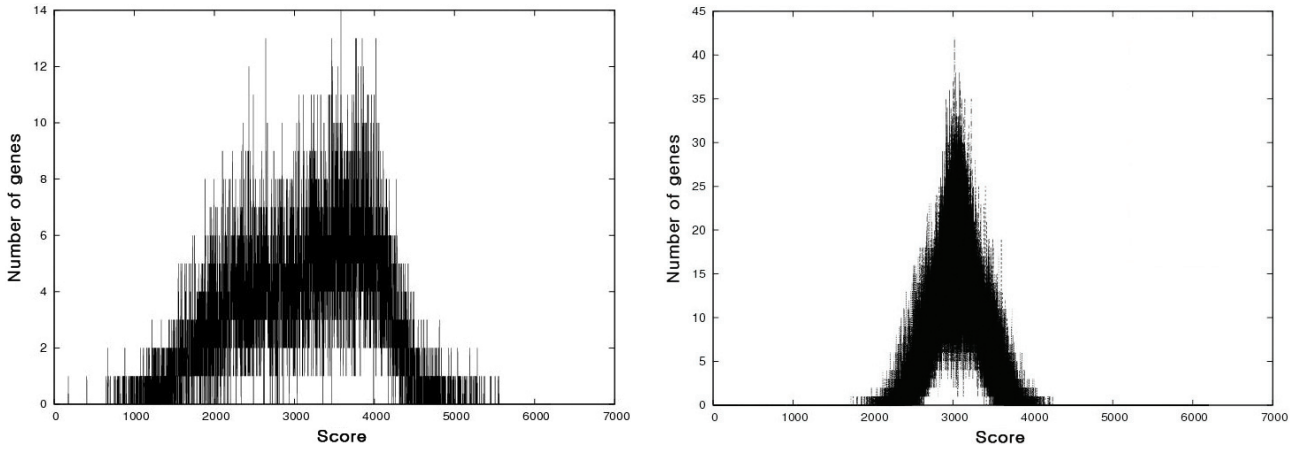


Fig. 4 Comparison of gene score distributions using raw and rank data (*Golub*'s data was used).

ted along the y-axis. Figure 4 (a) shows the score distribution from the *raw data*, where the score of each gene is calculated using the gene expression values of the samples. Figure 4 (b) shows the score distribution from the *rank data* where the score of each gene is calculated using the rank values of the samples. The results show that the two score distributions have very similar shapes and, as expected, show heavier tails, indicating that many genes are differentially expressed in the two classes. Figure 4 (c) shows a set of score distributions from the column-permuted *raw data*, and Fig. 4 (d) shows a set of score distributions from the column-permuted *rank data*. Here, only twenty score distributions are plotted for each case. The results show that the score distributions from the *raw/rank data* are more spread out with heavy tails, while the score distributions from the column-permuted *raw/rank data* are relatively concentrated with smaller variances. Based on the S_i values, p = 0.0005 was determined from the *rank data*. Also, p = 0.0053 was determined from the *raw data*, which is consistent with the p-value reported by Park [3]. This result verifies our expectation that the scoring result from the *rank data* is statistically significant.

To investigate whether more significant genes can be selected from integrated rank-based data, the scoring method was applied to the individual and integrated rank-based data. Note that the scale of the expression values of individual raw microarray data is quite different, and thus multiple raw microarray data cannot be integrated directly as explained in Sect. 3.1. Therefore, the gene scoring method cannot be applied to the integrated raw data.

Here, we used the Prostate Cancer data set. We first performed the permutation tests for three individual raw/rank microarray data: LaTulippe, Welsh, and Singh. Next, we performed the same test for each of four integrated rank-based microarray data: (LaTulippe+Welsh), (LaTulippe+Singh), (Welsh+Singh), and (LaTulippe+Welsh+Singh). Here, (LaTulippe+Welsh) represents the integrated data resulting from the merging of the LaTulippe and Welsh data. We performed 10,000 permutations for each experiment. We show the score distributions of the integrated rank-based data, (LaTulippe+Welsh+Singh), in Fig. 5. Figure 5 (a) shows the score distribution from the original integrated rank-based data, and Fig. 5 (b) shows a set of score distributions from the column-permuted integrated rank-based data. Here, only twenty score distributions are plotted in Fig. 5 (b).



(a) `integrated rank-based data` (b) `column-permuted integrated rank-based data`

Fig. 5 Gene score distributions of integrated rank-based data (LaTulipp+Welsh+Singh data was used).

Table 1 P-values obtained from individual and integrated data.

	raw data	rank data
LaTulippe	0.0037	0.0023
Welsh	0.0001	0.0001
Singh	0.0271	0.0414
LaTulippe + Welsh	NA	0.0269
LaTulippe + Singh	NA	0.0002
Welsh + Singh	NA	0.0031
LaTulippe + Welsh + Singh	NA	0.0001

Table 1 summarizes the experimental result of each permutation test. We obtained p-values of 0.0023, 0.0001, and 0.0414 for the three rank-based data, LaTulippe, Welsh, and Singh, respectively. We also obtained p-values of 0.0269, 0.0002, 0.0031, and 0.0001 for the four integrated rank-based data, (LaTulippe+Welsh), (LaTulippe+Singh), (Welsh+Singh), and (LaTulippe+Welsh+Singh), respectively. It is clear that the p-values obtained from each instance of individual and integrated rank-based data are less than the significance level of 0.05. Note that p-values are not available for the four integrated raw data. These results show that when we integrate the individual rank-based microarray data with a statistically significant scoring result, we can obtain integrated rank-based data whose scoring result is also statistically significant.

4.3 Evaluation of the Selected Informative Genes

A simple and standardized method is still unknown to evaluate the performance of different experimental and analytical methodologies for extracting informative genes from microarray data. In this study, a combination of two approaches was employed to enable cross-analysis of the informative genes obtained from either individual or integrated experimental data.

Firstly, the number of well-known marker genes (included in a set of informative genes extracted from each experiment data) was quantitatively compared and evaluated. The Cancer Genome Anatomy Project (CGAP) database was chosen as the reference database to search for marker genes.

Secondly, functional analyses were performed for the sets of informative genes extracted from each experiment data. We used information available from the GO database [5]. The GO database contains functional gene annotations in a hierarchical structure that reflect the relationship between the biological terms and associated genes. Within an ontology, the terms form a directed acyclic graph in which terms are children of one or several more general terms. The GO terms do not describe specific gene or gene products themselves. Rather, the collaborating database generates gene association information consisting of links between genes or gene products and GO terms [28].

4.3.1 Comparison of Informative Genes Extracted from Single and Integrated Data

In this section, a comparative analysis is conducted between informative genes extracted from a single microarray data and those extracted from integrated rank-based microarray data.

A list of cancer marker genes registered in the CGAP database was utilized in order to compare the informative genes extracted from each experimental data. For each gene, the CGAP (Gene Info) contains data on the gene expression level (being up/down-regulated) in a specific cancer or normal cell. It also includes the statistical significance of the observed data (normal vs. a specific cancer) given the null hypothesis that the gene is expressed equally in normal and a specific cancer. Here, the statistical analysis is based on the Fisher exact test [4]. We used the Prostate Cancer data set as the experimental data, and compared the number of prostate cancer marker genes included in the sets of informative genes extracted from each experimental data. Following the retrieval of each gene from the CGAP (Gene Info), the gene was identified as a marker gene of prostate cancer only

if the value of statistical significance was less than 0.05 for the prostate tissue.

The first experiment involved extracting the 50 most differentially regulated informative genes from each experimental data and conducting a comparative evaluation of them. Of the 50 informative genes, 25 genes were up-regulated and the others were down-regulated. Three microarray data, LaTulippe, Welsh, and Singh, were used for the individual analysis, and an integrated rank-based data (LaTulippe+Welsh+Singh) was used for the analysis of combined, integrated data. A list of informative genes extracted from these experiment data is summarized in Appendix A. U1–U25 and D1–D25 show up- and down-regulated genes by rank, respectively.

Next, the information of each extracted gene was retrieved from the CGAP Gene Info, and the value of statistical significance, which informs as to whether a gene is a marker for prostate cancer, was extracted. In Appendix A, genes with p-values of less than 0.05 are shown in bold. The comparison results of this table can be briefly summarized as follows: LaTulippe contained 10 marker genes (20% of the 50 genes) of which 5 were up-regulated genes and the other 5 were from the list of down-regulated genes. Welsh included 19 marker genes (38%) of which 8 were in the list of up-regulated genes and the rest were down-regulated genes. Singh was most numerous among the individual data, and included 21 marker genes (42%) of which 11 were in the list of up-regulated genes and the rest were in the list of down-regulated genes. This could be due to the fact that Singh had more samples than LaTulippe or Welsh, as previously mentioned. For the integrated data, (LaTulippe+Welsh+Singh) included the most marker genes, a total of 26 marker genes (52%) of which 15 were up-regulated genes and the other 11 were down-regulated genes.

Figure 6 is a representative graph in which the number of marker genes extracted from each experiment is compared and indicated. As shown in this figure, the informative genes extracted from the integrated data generally include more marker genes for prostate cancer compared to those extracted from a single data, providing evidence that more informative genes can be extracted through integrated data.

As a second experiment, we conducted a functional analysis of informative genes using GO [5]. The functional attributes of informative genes are expected to reflect biological associations between genes and cancer disease. The informative genes corresponding to the top and bottom 2.5% of the ranked genes were extracted from each single and integrated microarray data, respectively. The relevant terms for each informative gene were subsequently extracted from the GO database. To identify statistically significant GO terms, a χ^2 (chi-square) test [17] was used and a p-value for each term was calculated. We concentrated on those terms significant at 5% (p-value < 0.05). From these experiments, 56, 57, and 30 terms were selected from LaTulippe, Welsh, and Singh, respectively, and 51 terms were selected from the integrated data (LaTulippe+Welsh+Singh).

To enhance the interpretation of such a list of terms, we categorized these terms into the following three groups: 1. Cell proliferation/protein synthesis; 2. Cell differentiation; and 3. Cell migration/invasion/angiogenesis. Cancer may be defined as unregulated, continuous cell proliferation. The terms categorized into the first group for the cell cycle, cell proliferation and protein synthesis, are therefore considered directly related to cancer disease [29]. In addition, cell differentiation is closely related to the control of cell proliferation. Cancer is also characterized by a change in the status of cell growth, as the development of a fertilized egg or stem cells into multi-functional somatic cells frequently involves repeated cell growth and differentiation [30]. Lastly, some terms are related to cancer progression, which makes a decisive contribution to the malignancy of the cancer cells, not directly related to non-cancerous cells' metastasis into cancer cells. Cell proliferation is not enough for cancerous cells to become malignant. This requires concrete changes such as migration, invasion, and angiogenesis [31].

Table 2 shows the results of the functional categorization of the related terms extracted from each single

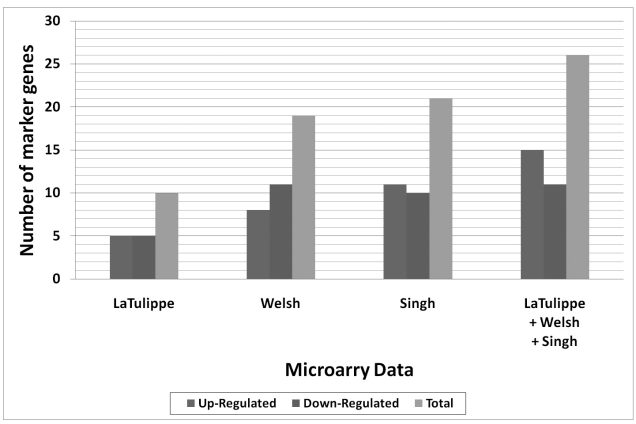


Fig. 6 Comparison of the number of marker genes extracted from each individual and integrated data (For each data set, the 50 most differentially regulated genes were evaluated).

Table 2 Comparison of the number of related terms which belong to each functional category (For each data set, the top 5% most differentially regulated genes were evaluated).

functional category \ data	LaTulippe	Welsh	Singh	LaTulippe + Welsh + Singh
Cell Proliferation/ Protein Synthesis	13/56	10/57	11/30	18/51
	23.2%	17.5%	36.7%	35.3%
Cell Differentiation	12/56	5/57	3/30	8/51
	21.4%	8.8%	10%	15.7%
Migration/ Invasion/ Angiogenesis	7/56	10/57	1/30	6/51
	12.5%	17.5%	3.3%	11.8%
Total	27/56	25/57	15/30	32/51
	48.2%	43.8%	50%	62.7%

and integrated microarray data. In the case of LaTulippe, the number of terms with significance probability of < 0.05 of significance probability was 56, where 13 were related to cell proliferation/protein synthesis. Numerous terms were found for cell proliferation/protein synthesis as well as for cell differentiation in LaTulippe and for migration/invasion/angiogenesis in Welsh. As explained earlier, these three categories are closely related to cancer diagnosis. In the case of (LaTulippe+Welsh+Singh), more terms were found evenly throughout the three categories compared to the single experimental data. This indicates that the integrated microarray experiment is more efficient in functional analyses of genes for cancer diagnosis, compared to a single-microarray experiment.

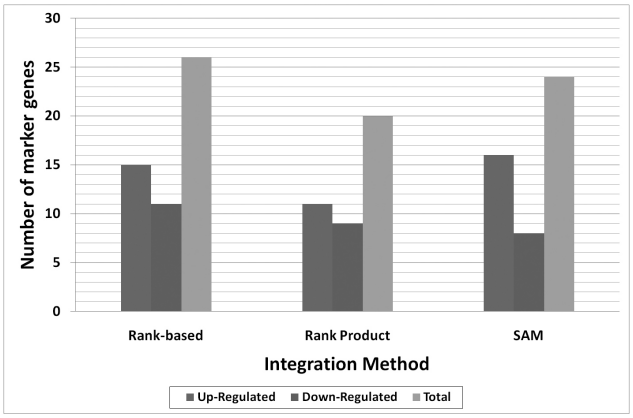


Fig. 7 Comparison of the number of marker genes extracted from each integrated data (For each data set, the 50 most differentially regulated genes were evaluated).

4.3.2 Comparison of Three Microarray Data Integration Methods

In this section, the rank-based method is compared with other microarray data integration methods. For this comparison, the following two methods are used. The first method integrates microarray data according to the z-score-based normalization and transformation technique [11] and extracts informative genes by SAM (Significance Analysis of Microarrays) [32] based on t-statistics. The second method is the Rank Product model [9], which is a meta-analysis method.

The Prostate Cancer data set was also used as the experimental data for this comparison. We first merged the three individual data, LaTulippe, Welsh, and Singh, according to the three respective methods of the rank-based method, the normalization and transformation method, and the Rank Product method, resulting in three integrated microarray data. We refer to the integrated microarray data that are merged by each method as Rank-based, SAM, and Rank Product, respectively, in this paper. We then compared the number of prostate cancer marker genes that were included in the sets of informative genes extracted from each of the three integrated data. As explained in Sect. 4.3.1, we used the list of marker genes for prostate cancer, which is included in the CGAP database.

As in the first experiment, the 50 most differentially expressed genes from each integrated data were extracted and comparatively evaluated. 25 genes were selected as up- and down-regulated informative genes, respectively. Appendix B shows a list of the informative genes extracted from each integrated data. U1–U25 and D1–D25 show up- and down-regulated genes, respectively.

Next, the information of each extracted gene was retrieved from the CGAP Gene Info, and the gene was identified as a marker gene of prostate cancer if the value of statistical significance was less than 0.05. In Appendix B, genes with p-values of less than 0.05 are shown in bold.

The experimental results shown in Appendix B can be briefly summarized as follows: Rank Product includes 20 marker genes (40% of the 50 differentially regulated genes) of which 11 are up-regulated genes and 9 are down-regulated genes. SAM contains 24 marker genes (48% of the 50 genes) of which 16 are up-regulated genes and 8 are down-regulated genes. Rank-based includes 26 marker genes (52% of the 50 genes) of which 15 are up-regulated genes and 11 are down-regulated genes.

Figure 7 is a representative graph in which the number of marker genes is compared and indicated. As shown in this figure, the informative genes extracted from the Rank-based generally include more marker genes for prostate cancer compared to those extracted from Rank Product or SAM. This provides evidence that more informative genes can be extracted from the rank-based integrated data.

To compare these lists of informative genes extracted from each integrated data, we also conducted the following functional analysis using the procedure explained in Sect. 4.3.1. The informative genes corresponding to the up- and down-regulated 2.5% of the ranked genes, were extracted from the Rank-based, SAM, and Rank Product, respectively.

Next, the relevant terms for each informative gene were extracted from the GO database and statistically significant GO terms (p-value < 0.05) were selected for comparison. From these experiments, 51, 48, and 46 terms were extracted from the Rank-based, SAM, and Rank Product, respectively.

The selected terms were then classified into the three categories of Cell Proliferation/protein synthesis, Cell differentiation, and Cell migration/invasion/angiogenesis. As explained in Sect. 4.3.1, these three categories are closely related to cancer diagnosis. Table 3 shows the comparison results of the functional categorization of the related terms extracted from each integrated data. In Table 3, the ratio represents how many terms were classified into each category among all related terms. The corresponding percentage is also given in the second column. In the case of Rank-based, more terms were found evenly throughout the three categories compared to the other experimental data. This reflects that more cancer-related terms were extracted from the Rank-based data. This experiment serves as evidence that our rank-based approach provides more efficient functional

Table 3 Comparison of the number of related terms which belong to each functional category (For each data set, the top 5% most differentially regulated genes were evaluated).

data / functional category	Rank-based	Rank Product	SAM
Cell Proliferation/ Protein Synthesis	18/51	10/48	17/46
	35.3%	20.8%	37%
Cell Differentiation	8/51	6/48	5/46
	15.7%	12.5%	10.9%
Migration/ Invasion/ Angiogenesis	6/51	3/48	1/46
	11.8%	6.3%	2.2%
Total	32/51	19/48	23/46
	62.7%	39.6%	50%

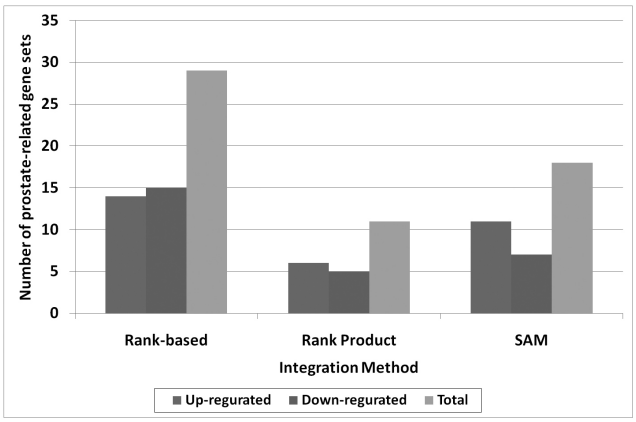


Fig. 8 Comparison of the number of prostate-related gene sets extracted from each integrated data (For each data set, the top scoring 40 gene sets were evaluated).

profiling for cancer diagnosis compared to other microarray integration methods.

In addition, we evaluated each of three integrated data by the GSEA method [6], [7], and compared the differentially regulated gene sets. In reference [6], Subramanian et al. proposed a method called GSEA for assessing the significance of pre-defined gene sets, rather than individual genes. GSEA determines whether members of a gene set tend to occur toward the top (or bottom) of the ranked genes list, in which case the gene set is correlated with the phenotypic class distinction. In GSEA, enrichment score (ES) reflects the degree to which a set is overrepresented at the extreme (top or bottom) of the entire ranked genes list.

To identify significant gene sets correlated with prostate cancer, we performed GSEA on the Rank-based, SAM, and Rank Product data with the CM (Cancer Modules) catalog of cancer-related gene sets, and analyzed the results which consist of the enriched gene sets with ES and FDR. Specifically, we considered the top scoring 40 gene sets (FDR $\leq$ 0.25) in each of the three integrated data and their corresponding clinical annotations to better understand the underlying biology in the experimental data. 20 gene sets were selected as up- and down-regulated gene sets, respectively. We identified an enriched gene set as the prostate-related one only if the gene set was annotated with the clinical attribute (tumor type) of prostate cancer with p-value < 0.05.

The experimental results can be briefly summarized as follows: Rank product included 11 prostate-related gene sets (27% of the 40 differentially regulated gene sets) of which 6 were up-regulated and 5 were down-regulated. SAM contained 18 prostate-related gene sets (45% of the 40 gene sets) of which 11 were up-regulated and 7 were down-regulated. Rank-based approach included 29 prostate-related gene sets (72% of the 40 gene sets) of which 14 were up-regulated and 15 were down-regulated. Figure 8 compares the number of prostate-related gene sets extracted from each of the three integrated data. As shown in this figure, the Rank-based included more enriched gene sets correlated with prostate cancer compared to those extracted from Rank Product or SAM. The result indicates that more informative gene sets can be extracted from the rank-based integrated data.

5. Conclusion

In this paper, we verified the usefulness of rank-based data integration and analysis methods. First, we showed that a non-parametric technique can be employed effectively as a scoring metric for rank-based microarray data and that the scoring results from integrated rank-based microarray data are statistically significant. Next, we verified experimentally that informative genes can be selected from integrated rank-based data. We used public prostate cancer microarray data for the experiments, the representative genes list from the CGAP database for the assessments and comparisons of the extracted genes, and the GO database and the GSEA method for the understanding of the functional associations between the extracted genes and cancer-related activities.

By comparing the informative genes extracted from individual microarray data with those extracted from integrated rank-based microarray data, we showed that the informative genes from integrated rank-based microarray data are statistically more significant. Moreover, by comparing the informative genes obtained by the rank-based approach, the z-score-based normalization technique, and the rank products technique, we verified that the rank-based data integration method enables the extraction of more informative genes related to cancer outbreaks and diagnoses.

The rank-based approach verified in this paper can integrate two or more microarray data by iteratively executing the merge step of the algorithm in Sect. 3.2. In addition, through the process of transforming raw data into rank data, it becomes possible to extract informative genes more accurately and robustly from multiple microarray data even with different platforms (e.g., cDNA or Affymetrix) or various experimental settings.

As a future research, we are planning to discover the regulatory network among the informative genes extracted

from the integrated rank-based microarray data, and extend the proposed rank-based integration method to the environments in which there are more than two class labels (i.e. stages of tumor progression).

Acknowledgement

This research was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (No. R01-2006-000-11106-0).

References

[1] Y.M. Yoon, J.C. Lee, and S.H. Park, "Building a classifier for integrated microarray datasets through two-stage approach," Proc. IEEE Symposium on Bioinformatics & Bioengineering, vol.6, pp.94–102, 2006.

[2] Y.M. Yoon, J.C. Lee, S.H. Park, S.J. Bien, H.C. Chung, and S.Y. Rha, "Direct integration of microarrays for selecting informative genes and phenotype classification," Inf. Sci., vol.178, pp.88–105, 2008.

[3] P.J. Park, M. Pagano, and M. Bonetti, "A nonparametric scoring algorithm for identifying informative genes from microarray data," Pacific Symposium on Biocomputing, vol.6, pp.52–63, 2001.

[4] L.H. Grouse, P.J. Munson, and P.S. Nelson, "Sequence databases and microarrays as tools for identifying prostate cancer biomarkers," Urology, vol.57, pp.154–159, 2001.

[5] The Gene Ontology Consortium, "Gene ontology: Tool for the unification of biology," Nature Genetics, vol.25, pp.25–29, 2000.

[6] A. Subramanian, et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," PNAS, vol.102, no.43, pp.15545–15550, 2005.

[7] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," Annals of Applied Statistics, vol.1, no.1, pp.107–129, 2007.

[8] D.R. Rhodes, T.R. Barrette, M.A. Rubin, D. Ghosh, and A.M. Chinnaiyan, "Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," Cancer Research, vol.62, pp.4427–4433, 2002.

[9] R. Breitling, P. Armengaud, and P. Herzyk, "Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments," Federation of European Biochemical Societies Letters, vol.573, pp.83–92, 2004.

[10] H. Jiang, et al., "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," BMC Bioinformatics, vol.5, pp.81–93, 2004.

[11] C. Cheadle, M.P. Vawter, W.J. Freed, and K.G. Becker, "Analysis of microarray data using Z score transformation," J. Molecular Diagnostics, vol.5, no.2, pp.73–81, 2003.

[12] L. Xu, A.C. Tan, D.Q. Naiman, D. Geman, and R.L. Winslow, "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data," Bioinformatics Advance Access, vol.21, no.20, pp.3905–3911, 2005.

[13] P. Warnat, R. Eils, and B. Brors, "Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes," BMC Bioinformatics, vol.6, p.265, 2005.

[14] H. Xiong, Y. Zhang, X. Chen, and J. Yu, "Cross-platform microarray data integration using the normalised linear transform," International Journal of Data Mining and Bioinformatics, vol.4, pp.142–157, 2010.

[15] K.L. Lange, R.J.A. Little, and J.M.G. Taylor, "Robust statistical modeling using the t distribution," J. American Statistical Association, vol.84, pp.881–896, 1989.

[16] H. Liu, C. Chen, Y. Liu, C. Chu, D. Liang, L. Shih, and C. Lin, "Cross-generation and cross-laboratory predictions of Affymetrix microarrays by rank-based methods," J. Biomedical Informatics, vol.41, no.4, pp.570–579, Aug. 2008.

[17] S. Drăghici, P. Khatri, R.P. Martins, G.C. Ostermeier, and S.A. Krawetz, "Global functional profiling of gene expression," Genomics, vol.81, pp.98–104, 2003.

[18] T.R. Golub, et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," Science, vol.286, pp.531–537, 1999.

[19] S. Rogers, R.D. Williams, and C. Campbell, "Class prediction with microarray datasets in: Bioinformatics using computational intelligence paradigms," Studies in Fuzziness and Soft Computing, vol.176, pp.119–141, 2005.

[20] L. Deng, J. Pei, J. Ma, and D.L. Lee, "A rank sum test method for informative gene discovery," Proc. 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol.176, pp.410–419, 2004.

[21] I.H. Witten and E. Frank, DATA MINING Practical Machine Learning Tools and Techniques, pp.97–112, Morgan Kaufmann, San Francisco, 2005.

[22] R. Marko and K. Igor, "Theoretical and empirical analysis of ReliefF and RReliefF," Mach. Learn., vol.53, pp.23–69, 2003.

[23] B.S. Kim, J.Y. Song, K.Y. Wang, S.S. Lee, and D.J. Chung, "Prostate cancer classification processor using DNA computing technique," IEICE Electronics Express, vol.6, no.10, pp.581–586, 2009.

[24] N. Bailey, Statistical methods in biology, Cambridge University Press, Cambridge, 1995.

[25] E. LaTulippe, et al., "Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease," Cancer Research, vol.62, pp.4499–4506, 2002.

[26] J.B. Welsh, et al., "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer," Cancer Research, vol.61, pp.5974–5978, 2001.

[27] D. Singh, et al., "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol.1, pp.203–209, 2002.

[28] Z.X. Yin and J.H. Chiang, "Novel algorithm for coexpression detection in time-varying microarray data sets," IEEE/ACM Trans. Computing Biology and Bioinformatics, vol.5, pp.120–135, 2008.

[29] H. Douglas and A.W. Robert, "The hallmarks of cancer," Cell, vol.100, pp.57–70, 2000.

[30] K. Vermeulen, D.R.V. Bockstaele, and Z.N. Berneman, "The cell cycle: A review of regulation, deregulation and therapeutic targets in cancer," Cell Proliferation, vol.36, pp.131–149, 2003.

[31] S. Saadoun, W.C. Papadopoulos, M. Hara-Chikuma, and A.S. Verkman, "Impairment of angiogenesis and cell migration by targeted aquaporin-1 gene disruption," Nature, vol.434, pp.786–792, 2005.

[32] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," PNAS, vol.98, no.9, pp.5116–5121, 2001.

Appendix A: Comparison of Informative Genes Extracted from Single and Integrated Data. U1–U25 and D1–D25 Show Up- and Down-Regulated Genes, Respectively. Genes with p-Values of Less than 0.05 Are Shown in Bold.

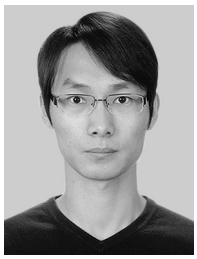
Rank	LaTulippe		Welsh		Singh		LaTulippe + Welsh + Singh	
	Gene Symbol	P-value	Gene Symbol	P-value	Gene Symbol	P-value	Gene Symbol	P-value
U1	**TPRC**	**0.02**	CXADR	0.42	**HPN**	**0.02**	**HPN**	**0.02**
U2	CELP	-	**HPN**	**0.02**	**TARP**	**0.02**	**HSPD1**	**0.00**
U3	GIF	-	LSR	0.42	**HSPD1**	**0.00**	**TARP**	**0.02**
U4	LMX1B	-	**PAICS**	**0.00**	**KLK3**	**0.00**	SIM2	0.20
U5	SLC6A5	-	PSMB7	0.40	PLA2G7	0.35	TACSTD1	0.24
U6	ZBTB25	-	CANX	0.09	**PDLIM5**	**0.02**	C7orf24	0.29
U7	AMH	0.47	**SEC23B**	**0.00**	TMSL8	0.23	**PDLIM5**	**0.02**
U8	LOC90925	-	**KIAA0152**	**0.03**	XBP1	0.29	LOC728900	0.44
U9	AVP	-	**KRT18**	**0.00**	RGS10	0.23	**NME1**	**0.00**
U10	DHRS2	0.23	PPIB	0.39	GUCY1A3	0.21	TMSL8	0.23
U11	RBM19	0.42	LOC728900	0.44	NME2	-	**TSPAN1**	**0.02**
U12	HOXC6	-	TCEB2	0.32	THBS4	0.05	**MARCKSL1**	**0.00**
U13	ITGBL1	-	TACSTD1	0.24	-	0.31	FBP1	0.16
U14	C7orf24	0.29	C7orf24	0.29	**LOC284821**	**0.00**	BOLA2	0.47
U15	**AMACR**	**0.00**	PCCB	0.17	**SLC25A6**	**0.00**	**PYCR1**	**0.00**
U16	DEFA1	-	IQGAP2	0.18	SIM2	0.20	PDIA5	0.29
U17	SPAG6	-	TMEM4	0.15	**TSPAN1**	**0.02**	**IMPDH2**	**0.00**
U18	**PRPF19**	**0.00**	**CCT3**	**0.00**	C7orf24	0.29	**BRP44**	**0.03**
U19	**LLGL2**	**0.02**	TMED3	0.33	FBP1	0.16	**RPS2**	**0.00**
U20	GNL2	0.31	ANAPC5	0.49	TACSTD1	0.24	**KRT18**	**0.00**
U21	ALCAM	0.23	DEAF1	0.44	**LOC643779**	**0.00**	**FASN**	**0.00**
U22	MMP19	0.23	**MRPL12**	**0.01**	LOC728900	0.44	**RPS18**	**0.00**
U23	**COMP**	**0.02**	BICD1	0.20	AGR2	0.25	RAP1GAP	0.10
U24	DIAPH1	0.32	BBS4	0.35	**FOLH1**	**0.00**	-	0.31
U25	EPRS	0.05	**NME1**	**0.00**	**MARCKSL1**	**0.00**	**UAP1**	**0.01**
D1	**FCGRT**	**0.02**	**RBPMS**	**0.00**	**PTGDS**	**0.00**	STAC	0.15
D2	SOX5	-	**MYL6**	**0.00**	NELL2	0.15	**PTGDS**	**0.00**
D3	**LCAT**	**0.00**	CLU	0.06	ANXA2	–	ANXA2	–
D4	PNMT	-	PSIP1	0.24	**SPON1**	**0.04**	RCAN2	0.23
D5	IGF2	-	ANGPT1	0.16	STAC	0.15	ANGPT1	0.16
D6	CYP3A43	-	RCAN2	0.23	LMO3	0.35	PRKCA	0.10
D7	MDM1	0.42	FZD7	0.34	**CFD**	**0.00**	**CRYAB**	**0.00**
D8	ERC1	0.20	CBX7	0.31	CALM1	0.36	**COL4A6**	**0.02**
D9	COL13A1	0.38	**KLHL21**	**0.00**	**COL4A6**	**0.02**	**HEPH**	**0.01**
D10	ANGPT1	0.16	FTO	0.15	PRKCA	0.10	**SPON1**	**0.04**
D11	CHRFAM7A	-	DMPK	0.08	HOM-TES-103	0.09	ITSN1	0.38
D12	LDOC1	0.23	RRAS	0.24	ITSN1	0.38	IL11RA	0.06
D13	**GPR161**	**0.02**	**LAPTM4A**	**0.02**	EPB41L3	0.35	-	0.15
D14	**CX3CR1**	**0.03**	**TRIP6**	**0.02**	**CDC42BPA**	**0.04**	LMO3	0.35
D15	KIAA0888	0.47	-	0.15	**FAM107A**	**0.01**	VCL	0.42
D16	-	0.15	PPP3CB	0.36	**PENK**	**0.01**	NELL2	0.15
D17	ATRNL1	0.15	**SVIL**	**0.01**	RCAN2	0.23	CALM1	0.36
D18	**SPEG**	**0.01**	SRF	0.34	**FLRT2**	**0.04**	**TP63**	**0.01**
D19	IGSF1	0.23	**DES**	**0.00**	GSTM1	0.42	**FLRT2**	**0.04**
D20	MAGI2	0.32	**PPP1R12B**	**0.00**	IL11RA	0.06	DIP2C	0.39
D21	MKLN1		**PALLD**	**0.00**	ANGPT1	0.16	**AOX1**	**0.04**
D22	RP4-691N24.1	0.35	OPTN	0.42	CLEC3B	0.23	**STOM**	**0.01**
D23	TRO	0.05	**FLNA**	**0.00**	**CRYAB**	**0.00**	**PENK**	**0.01**
D24	LYST	0.20	DST	0.07	AKR1B1	0.38	**MYL6**	**0.00**
D25	KPNA3	0.32	**MYLK**	**0.00**	**ABL1**	**0.03**	KCNMB1	0.09

Appendix B: Comparison of Informative Genes Extracted from Each Integrated Data. U1–U25 and D1–D25 Show Up- and Down-Regulated Genes, Respectively. Genes with p-Values of Less than 0.05 Are Shown in Bold.

Rank	Rank-based		Rank product		SAM	
	Gene Symbol	P-value	Gene Symbol	P-value	Gene Symbol	P-value
U1	**HPN**	**0.02**	**HPN**	**0.02**	**HPN**	**0.02**
U2	**HSPD1**	**0.00**	**AMACR**	**0.00**	**TARP**	**0.02**
U3	**TARP**	**0.02**	C7orf24	0.29	**PDLIM5**	**0.02**
U4	SIM2	0.20	TACSTD1	0.24	TACSTD1	0.24
U5	TACSTD1	0.24	**HSPD1**	**0.00**	**TSPAN1**	**0.02**
U6	C7orf24	0.29	**PDLIM5**	**0.02**	-	-
U7	**PDLIM5**	**0.02**	HOXC6	-	-	0.31
U8	LOC728900	0.44	**PYCR1**	**0.00**	FBP1	0.16
U9	**NME1**	**0.00**	RGS10	0.23	**MARCKSL1**	**0.00**
U10	TMSL8	0.23	**NME1**	**0.00**	TMSL8	0.23
U11	**TSPAN1**	**0.02**	BOLA2	0.47	RGS10	0.23
U12	**MARCKSL1**	**0.00**	DOPEY2	0.37	**KRT18**	**0.00**
U13	FBP1	0.16	WSB2	0.47	**PYCR1**	**0.00**
U14	BOLA2	0.47	NME1-NME2	-	**FASN**	**0.00**
U15	**PYCR1**	**0.00**	FBP1	0.16	**NME1**	**0.00**
U16	PDIA5	0.29	CLDN8	0.42	**BRP44**	**0.03**
U17	**IMPDH2**	**0.00**	LYPLA1	0.21	**GDF15**	**0.00**
U18	**BRP44**	**0.03**	**UAP1**	**0.01**	**IMPDH2**	**0.00**
U19	**RPS2**	**0.00**	-	0.31	CLDN3	0.15
U20	**KRT18**	**0.00**	SIM2	0.20	**HEBP2**	**0.00**
U21	**FASN**	**0.00**	**MYO6**	**0.00**	**FOXA1**	**0.00**
U22	**RPS18**	**0.00**	**KRT18**	**0.00**	**HSPD1**	**0.00**
U23	RAP1GAP	0.10	SIM2	0.20	BCAM	0.16
U24	-	0.31	**MYO6**	**0.00**	C7orf24	0.29
U25	**UAP1**	**0.01**	**KRT18**	**0.00**	**AMACR**	**0.00**
D1	STAC	0.15	ANGPT1	0.16	**PTGDS**	**0.00**
D2	**PTGDS**	**0.00**	LMO3	0.35	ANXA2	-
D3	ANXA2	-	CALM1	0.36	**CFD**	**0.00**
D4	RCAN2	0.23	STAC	0.15	ANGPT1	0.16
D5	ANGPT1	0.16	**COL4A6**	**0.02**	**STOM**	**0.01**
D6	PRKCA	0.10	**MEIS2**	**0.02**	VCL	0.42
D7	**CRYAB**	**0.00**	**CLIP3**	**0.00**	CLU	0.06
D8	**COL4A6**	**0.02**	CCND2	0.10	**CRYAB**	**0.00**
D9	**HEPH**	**0.01**	GSTM1	0.42	-	-
D10	**SPON1**	**0.04**	**GSTP1**	**0.04**	**PTGDS**	**0.00**
D11	ITSN1	0.38	-	0.15	GSTM1	0.42
D12	IL11RA	0.06	**SPON1**	**0.04**	SYNPO	0.23
D13	-	0.15	GPM6B	0.47	**USP11**	**0.01**
D14	LMO3	0.35	**CRYAB**	**0.00**	**ZFP36L1**	**0.02**
D15	VCL	0.42	ANXA2	-	GAS6	0.39
D16	NELL2	0.15	**TGFB1I1**	**0.01**	KCNMB1	0.09
D17	CALM1	0.36	CHRDL1	0.37	STAC	0.15
D18	**TP63**	**0.01**	ITGA1	-	CALM1	0.36
D19	**FLRT2**	**0.04**	ATP1A2	0.10	MEG3	0.07
D20	DIP2C	0.39	VCL	0.42	ITSN1	0.38
D21	**AOX1**	**0.04**	**CDC42EP3**	**0.00**	SERPINF1	0.15
D22	**STOM**	**0.01**	RCAN2	0.23	TNNT1	0.15
D23	**PENK**	**0.01**	PRKCA	0.10	LMO3	0.35
D24	**MYL6**	**0.00**	C16orf45	0.28	NELL2	0.15
D25	KCNMB1	0.09	**FAM107A**	**0.01**	**RBP1**	**0.00**

Dongwan Hong received the M.S. degree and the Ph.D. degree in Computer Science and Computer Engineering from Hallym University in 1998 and 2007. Currently he is a Research Professor of the Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine. His research interests include Bioinformatics and Databases.

Jeehee Yoon received the M.S. degree and the Ph.D. degree in Information Engineering from kyushu University, Japan in 1985 and 1988. Currently she is a Professor of Department of Computer Engineering, Hallym University. Her research interests include DNA Sequence search, Shape-based retrieval in time-series databases and Microarray data analysis.



Jongkeun Lee is a Ph.D. student at Department of Computer Engineering, Hallym University, Korea. He received the B.S. and the M.S. degrees in Computer Engineering from Hallym University in 2005 and 2007. His research interests include Bioinformatics and Databases.

Sanghyun Park received the B.S. and the M.S. degrees in Computer Engineering from Seoul National University in 1989 and 1991. He received the Ph.D. degree in Computer Science from UCLA in 2001. Currently he is a professor of Department of Computer Science, Yonsei University. His research interests include Databases, Data mining and Bioinformatics.

Jongil Kim received the M.S. degree and the M.D. degree in Medical College from Seoul National University in 1992 and 1999. Currently he is a professor at Seoul National University. His research interests include Molecular Genetics and Genomic Medicine.