Genomics & Informatics Vol. 6(3) 136-141, September 2008

Disease Prediction Using Ranks of Gene Expressions

Ki-Yeol Kim[1], Dong Hyuk Ki[2,3], Hyun Cheol Chung[2,3,4] and Sun Young Rha[2,3,4]*

[1]Oral Cancer Research Institute, Yonsei University College of Dentistry, [2]Cancer Metastasis Research Center, [3]Brain Korea 21 Project for Medical Science, [4]Department of Internal Medicine, Yonsei University College of Medicine, Seoul 120-752, Korea

Abstract

A large number of studies have been performed to identify biomarkers that will allow efficient detection and determination of the precise status of a patient' s disease. The use of microarrays to assess biomarker status is expected to improve prediction accuracies, because a whole-genome approach is used. Despite their potential, however, patient samples can differ with respect to biomarker status when analyzed on different platforms, making it more difficult to make accurate predictions, because bias may exist between any two different experimental conditions. Because of this difficulty in experimental standardization of microarray data, it is currently difficult to utilize microarray-based gene sets in the clinic. To address this problem, we propose a method that predicts disease status using gene expression data that are transformed by their ranks, a concept that is easily applied to two datasets that are obtained using different experimental platforms. NCI and colon cancer datasets, which were assessed using both Affymetrix and cDNA microarray platforms, were used for method validation. Our results demonstrate that the proposed method is able to achieve good predictive performance for datasets that are obtained under different experimental conditions.

Keywords: biomarker, different platform, microarray, gene expression, rank, prediction

Introduction

To identify disease-specific genes, numerous datasets have been created under different experimental conditions at different laboratories, albeit for the same purpose. Many aspects of these results, which have been derived from different datasets, are inconsistent, even though the datasets were created with the same objective using the same or similar technical platforms. Therefore, methods to integrate the results from different datasets (Moreau, *et al.*, 2003; Rhodes, *et al.*, 2002) as well as methods to combine datasets prior to analysis have been studied (Jiang, *et al.*, 2004; Kim, *et al.*, 2007; Lee, *et al.*, 2004). Such studies have shown that reliable results can be obtained by integrating results that are derived from different datasets and analyzing the combined datasets, as long as an increasing number of samples are utilized. The use of categorized values of gene expression ratios may improve prediction accuracies with respect to classification of different experimental datasets (Huan *et al.*, 2002). One approach that has been suggested is the discretization of gene expression levels (George *et al.*, 2004). This method divides continuous gene expression levels into several categories, and, as a result, the bias that can exist between different microarray datasets is minimized.

Properly developed integration approaches should help identify biomarkers to classify specific diseases based on high-throughput data. However, when a patient' s sample is evaluated to determine his/her disease status using more than one experimental condition relative to a determined biomarker set, correct prediction becomes impossible. Furthermore, methods to predict the disease status of a patient using biomarkers that initially are identified under different conditions than those that are used for the patient analysis have not been developed.

This study suggests a method that can accurately predict the disease status of a patient using a predetermined biomarker that is developed on a different platform. Specifically, we performed a two-step discretization of gene expression values by their rank, which were processed in both the biomarker selection and prediction stages.

Methods

Datasets

To evaluate our proposed method, we used two different datasets: the NCI dataset (Lee, *et al.*, 2003) and the colon cancer dataset (Kim, *et al.*, 2007; Notterman, *et al.*, 2001). Both of these datasets include gene expression information that was determined experimentally using two different microarray platforms (oligonucleotide-based and cDNA-based). There are a large number

*Corresponding author: E-mail rha7655@yuhs.ac
Tel +82-2-2228-8050, Fax +82-2-362-5592
Accepted 20 August 2008

of cancer tissue types in the NCI dataset; however, we used only the expression data of ovarian cancer and colon cancer tissues in this study. The datasets that were used in this study are summarized in Table 1.

Selection of significant gene sets from the training dataset

For transformation of the dataset, gene expression ratios were ranked in order of expression ratio for each gene, and the ranks were matched with the corresponding experimental group. This process is similar to the first step of the nonparametric Mann-Whitney U test. The steps that were used to discretize gene expression levels are summarized below:

(1) Gene expression ratios were ranked for every gene in each dataset that had more than two experimental groups.
(2) The rank and assignment order based on gene expression were listed for corresponding experimental groups.
(3) The results of (2) were summarized in the form of a contingency table for each gene.
(4) The relationship between gene expression patterns and experimental groups for each gene was tested.
(5) The discriminative gene set was selected by measuring statistical significance.

Disease status prediction for a new patient using the selected discriminative gene set

The test dataset was predicted using the selected discriminative gene set as follows:

(1) Gene expression ratios were re-ranked within each experiment in the discriminative gene set.
(2) A predictor was created by using the re-ranked discriminative gene set in (1).
(3) Gene expression ratios of the test sample to be predicted were ranked.
(4) The prediction accuracy of the ranked test dataset was calculated using the predictor created in step (2).

The processes for significant gene selection and disease prediction (B and C) were summarized in Fig. 1.

Statistical analysis

After the gene expression ratios were summarized in the form of a contingency table for each gene, a nonparametric statistical method was applied to the datasets to test independency between gene expression patterns and experimental groups. The Kruskal-Wallis test and Fisher's exact test were used for continuous

Table 1. Summary of datasets

Data name	Experimental platform	# of genes	# of total samples	Group A	Group B
Colon cancer dataset				Colon normal	Colon tumor
Kim et al., 2007	cDNA	12,319	78	35	43
Notterman et al., 2001	Affymetrix HU6800	7,464	36	18	18
NCI dataset				Ovarian cancer	Colon cancer
Lee JK et al., 2003	cDNA	2,344	13	6	7
	Affymetrix HU6800	2,344	13	6	7

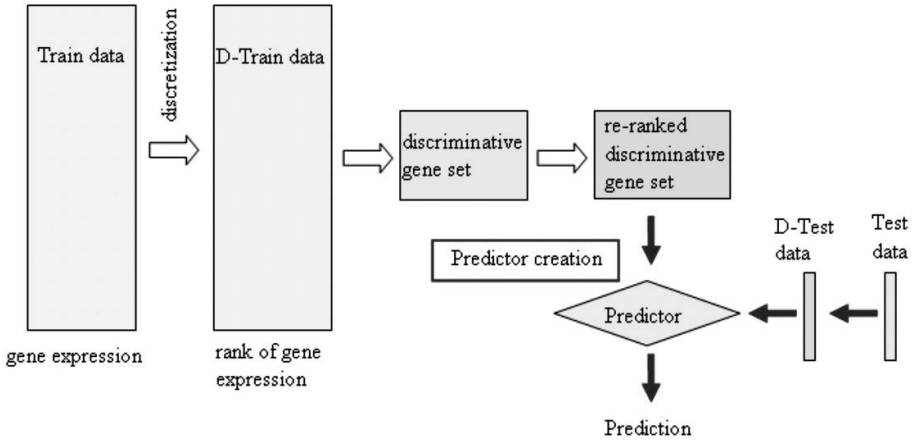

Fig. 1. Summary of the process for gene selection and prediction. D-Train data: discretized train dataset by rank; D-Test data: discretized test dataset by rank for each experiment.

gene expression and the discretized dataset, respectively. To evaluate the predictive accuracy of the selected significant gene set, the Random Forest (RF) test was used to enable re-sampling while still allowing for repetition (Breiman, 2001). We used the RF program in the R package (http://www.r-project.org) and calculated OOB (Out Of Bag) error and prediction accuracy as well.

We compared the prediction accuracies when real gene expression values were used and when the ranks of gene expression values were used. These two approaches can be summarized as follows.

(1) Method 1: Gene expression values were used for gene selection and prediction stages, and OOB error rate and prediction accuracy were calculated (OOB_raw data, Prediction_raw data).

(2) Method 2: During the gene selection and prediction stages, the ranks of gene expression were used (OOB_rank, Prediction_rank); this is the proposed method.

Results and Discussion

The OOB error rates and the prediction accuracies were compared for the two different approaches-namely, raw data versus rank data. In the NCI dataset, the prediction

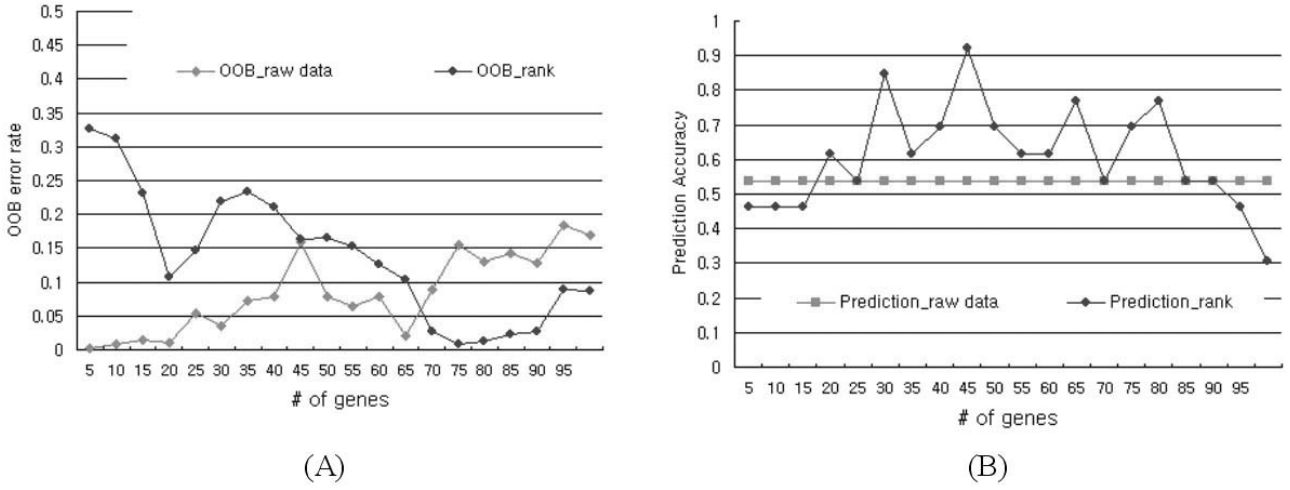

Fig. 2. Comparison of the OOB error rates (A) and prediction accuracies (B) for the NCI dataset. Oligo and cDNA datasets were used for training and testing, respectively.

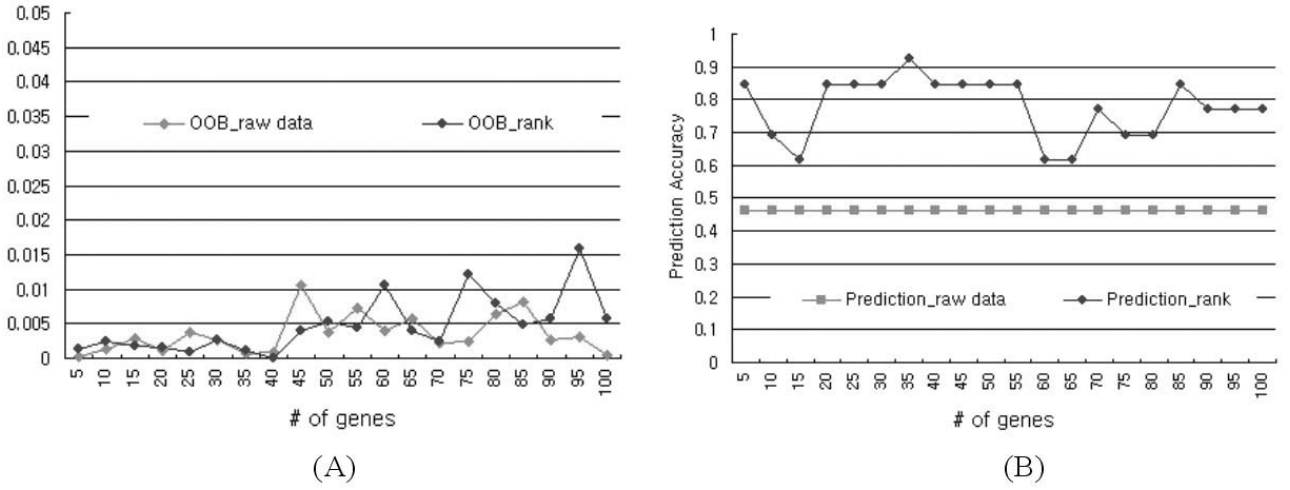

Fig. 3. Comparison of the OOB error rates (A) and prediction accuracies (B) for the NCI dataset. cDNA and Oligo datasets were used for training and testing, respectively.

accuracy was improved by using rank data (p=0.052275); however, the OOB error rate was significantly higher (p=0.037566) (Fig. 2).

We next compared the OOB error and prediction accuracy for the cDNA and Oligo datasets when they were used as training and testing sets, respectively. While the OOB error rate was not significantly different from the results in Fig. 3A (p=0.277), the prediction accuracy of the proposed method was significantly higher (p=4.57E-12) (Fig. 3B). In addition, the OOB error rate was almost 0 when the cDNA dataset was used as the training dataset (Fig. 3A). In contrast, the OOB error rates were high for both of the approaches when the Oligo dataset was used for training (Fig. 2A). Nevertheless, the prediction accuracies did not exhibit dependency on the OOB error rate (Fig. 2B, 3B).

In the NCI dataset, the fixed prediction accuracy was determined for the case when gene expression levels were used for both gene selection and prediction. This result indicated that all 13 tissues were classifiable into either the ovarian or colon cancer groups. By scaling the difference in gene expression between the training and testing datasets, the biomarker that was selected from the training dataset was not an accurate predictor of disease status when it was applied to the test dataset. In contrast, the prediction accuracy of the biomarker was improved when the discretized dataset was used for prediction. This result can be interpreted to

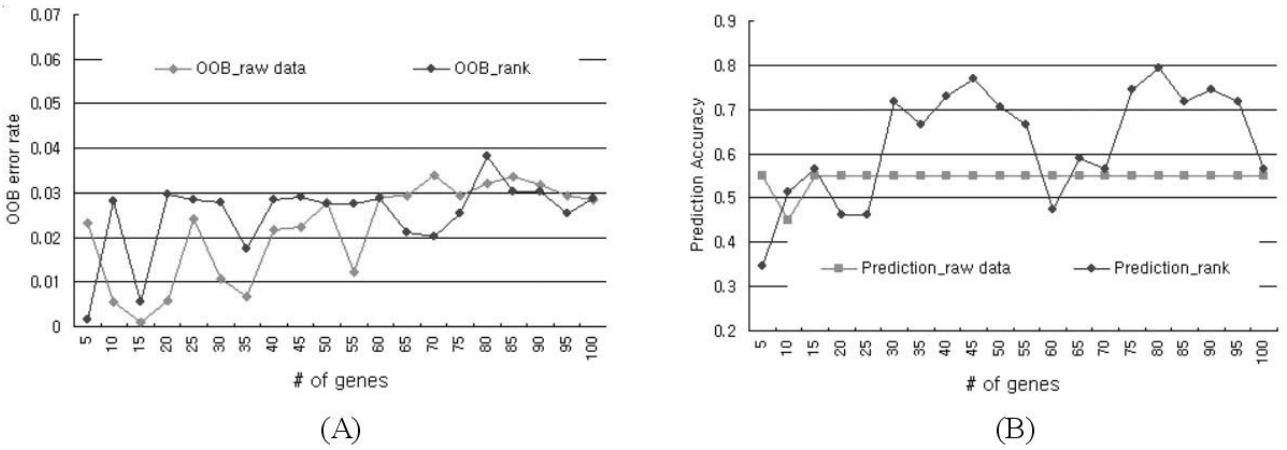

Fig. 4. Comparison of the OOB error rates (A) and prediction accuracies (B) for the colon dataset. Oligo and cDNA datasets were used for training and testing, respectively.

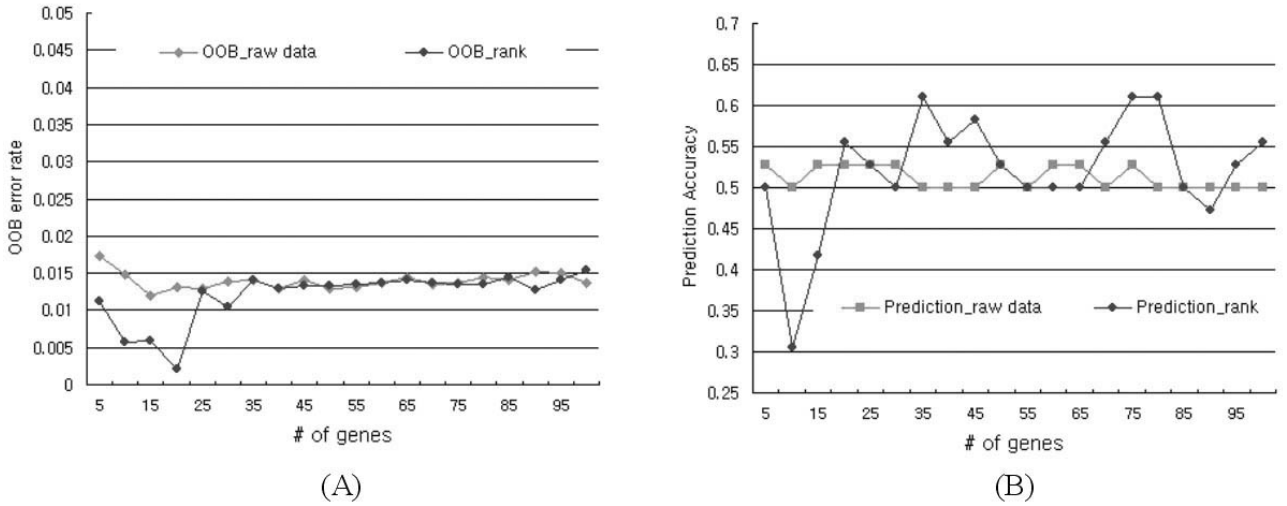

Fig. 5. Comparison of the OOB error rates (A) and prediction accuracies (B) for the colon dataset. cDNA and Oligo datasets were used for training and testing, respectively.

mean that the use of ranks compensated for the influence that different platforms had on prediction. Further, the prediction accuracy did not exhibit dependency on the number of significant genes. For this comparison, the significant genes were added one by one according to their level of significance. The fluctuation in prediction accuracies can be interpreted to mean that the added genes shuffled the order of gene expression and affected predictability. Our results clearly showed that the use of data that were discretized by rank was more effective when the cDNA dataset was used for training (Fig. 3) rather than when the Oligo dataset was used (Fig. 2).

For the colon cancer dataset, there was no significantly different OOB error rate between the two methods (p=0.33) (Fig. 4A). However, the prediction accuracy was significantly higher when the proposed method was used (p=0.0118) (Fig. 4B).

While the OOB error rate for the colon dataset was significantly different (p=0.0260) (Fig. 5A), the prediction accuracy was not (p=0.611) (Fig. 5B). As shown in Fig. 5A, the OOB error rate was significantly different by approximately 25 genes, becoming very similar afterward. Therefore, the significant OOB error rate was likely due to the differences that were observed within this range. However, the non-significance of the prediction accuracy could have been due to the very low prediction accuracies of the proposed method in this range (Fig. 5B).

The OOB error rate was slightly lower when the cDNA dataset was used as the training dataset; however, similar to what was observed for the NCI dataset, the prediction accuracy was not dependent on the OOB error rate. When the Oligo dataset was used as the training dataset and more than 25 to 30 significant genes were used, the proposed method exhibited good performance with respect to prediction accuracy (Fig. 4). However, there was some fluctuation in prediction accuracy when the cDNA dataset was used as the training dataset. Further, when the real gene expression levels were used, all tissues were classifiable into one of two classes, normal and tumor, which also was observed for the NCI dataset.

The use of biomarkers that are identified using microarrays can be expected to improve the prediction accuracies for a given disease status. However, when a sample from a patient is analyzed on a different experimental platform than that used to analyze the biomarker, it becomes difficult to make an accurate prediction of disease status, because bias can be generated when analyzing two different types of samples. Therefore, we developed a method to correctly predict disease status even when the new sample is analyzed on a different platform than that originally used to identify the biomarker.

In the NCI dataset, we found that it was effective to use the discretized value to select significant genes and make predictions; however, large variations in prediction accuracies were observed as the number of genes increased. This result likely was due to the fact that the NCI dataset contains a small number of samples, and thus misclassification of even one sample may impact the prediction accuracy. The prediction accuracy could be decreased by adding redundant genes when the number of genes is increased according to their significance.

Though the prediction accuracy of the proposed method was higher than the previous method that used continuous gene expression, it was not affected by the number of significant genes. Specifically, when the cDNA dataset was used as the training dataset, the OOB error rate was almost 0 and the selected gene set exhibited good predictive performance. This result indicates that the high and stable prediction accuracy was due to a reliably selected biomarker (Fig. 3).

In the colon cancer dataset, OOB error rates were low in both cases (Fig. 4A, 5A). While the prediction accuracy was improved by the proposed method when the Oligo dataset was used for training, it was not useful when the cDNA dataset was used for training.

For both datasets, the OOB error rates were lower and more stable when the cDNA dataset was used as the training dataset rather than the Oligo dataset. Further, if fewer than 50 significant genes were considered, the proposed method exhibited good performance for prediction of Oligo data with cDNA data, and vice versa.

For prediction of a new patient's disease status using biomarkers, the patient's sample should be analyzed using the same platform that was used to develop the biomarker to avoid bias and inaccurate results. However, this ideal situation is not always possible, and thus a reliable method to analyze data is needed for cases when the biomarker and patient sample have been processed on different platforms. The proposed method, which is capable of handling such an analysis, does so by transforming the gene expression values of a training dataset into discretized values and then selecting discriminative genes from the resulting discretized dataset.

Using the process that is outlined in this paper, the disease status of a patient can be predicted more reliably using a biomarker that is developed on a platform different to what was used to analyze the patient's sample. During the prediction stage, we transformed the selected significant genes by rank. By comparing the predictive accuracy of the number of significant genes,

we expect that a stably discriminative gene set that has a high predictive capacity can be produced.

Acknowledgments

This study was supported by a grant from the Korea Health 21 R&D Project, Ministry of Health & Welfare (0405-BC01-0604-0002) and the Korea Research Foundation (KRF-2005-005-J05904).

References

Jiang, H., Deng, Y., Chen, H.S., Tao, L., Sha, Q., Chen, J., Tsai, C.J., and Zhang, S. (2004). Joint analysis of two microarray gene-expression datasets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* 5, 81.

Kim, K.Y., Ki, D.H., Jeong, H.J., Jeung, H.C., Chung, H.C., and Rha, S.Y. (2007). Novel and simple transformation algorithm for combining microarray datasets. *BMC Bioinformatics* 8, 218.

Lee, J.K., Bussey, K.J., Gwadry, F.G., Reinhold, W., Riddick, G., Pelletier, S.L., Nishizuka, S., Szakacs, G., Annereau, J.P., Shankavaram, U., Lababidi, S., Smith, L.H., Gottesman, M.M., and Weinstein, J.N. (2003). Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells. *Genome Biol.* 4, R82.

Lee, J.S., Chu, I.S., Mikaelyan, A., Calvisi, D.F., Heo, J., Reddy, J.K., and Thorgeirsson, S.S. (2004). Application of comparative functional genomics to identify best-fit mouse models to study human cancer. *Nat. Genet* 36, 1306-1311.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., and Dabrowski, M. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet* 19, 570-577.

Notterman, D.A., Alon, U., Sierk, A.J., and Levine, A.J. (2001). Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res.* 61, 3124-3130.

Rhodes, D.R., Miller, J.C., Haab, B.B., and Furge, K.A. (2002). CIT: identification of differentially expressed clusters of genes from microarray data. *Bioinformatics* 18, 205-206.

Huan, L., Farhad, H., Chew, L.T., and Manoranjan, D. (2002). Discretization: an enabling technique, data. *Mining and Knowledge Discovery* 6, 393-423.

George, P., Lefteris, K., and Vassilis, M. (2004). Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. *3rd Hellenic Conference on Artificial Intelligence (SETN04).*

Breiman, L. (2001). Random Forests. Statistics Department, Berkeley, University of California, 1-33.