BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 13 2006, pages 1662–1664
doi:10.1093/bioinformatics/btl156

Gene expression

Modified signal-to-noise: a new simple and practical gene filtering approach based on the concept of projective adaptive resonance theory (PART) filtering method

Hiro Takahashi and Hiroyuki Honda*

Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan

Received on March 13, 2006; revised on April 6, 2006; accepted on April 19, 2006
Advance Access publication April 21, 2006
Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: Considering the recent advances in and the benefits of DNA microarray technologies, many gene filtering approaches have been employed for the diagnosis and prognosis of diseases. In our previous study, we developed a new filtering method, namely, the projective adaptive resonance theory (PART) filtering method. This method was effective in subclass discrimination. In the PART algorithm, the genes with a low variance in gene expression in either class, not both classes, were selected as important genes for modeling. Based on this concept, we developed novel simple filtering methods such as modified signal-to-noise (S2N′) in the present study. The discrimination model constructed using these methods showed higher accuracy with higher reproducibility as compared with many conventional filtering methods, including the *t*-test, S2N, NSC and SAM. The reproducibility of prediction was evaluated based on the correlation between the sets of *U*-test *p*-values on randomly divided datasets. With respect to leukemia, lymphoma and breast cancer, the correlation was high; a difference of >0.13 was obtained by the constructed model by using <50 genes selected by S2N′. Improvement was higher in the smaller genes and such higher correlation was observed when *t*-test, NSC and SAM were used. These results suggest that these modified methods, such as S2N′, have high potential to function as new methods for marker gene selection in cancer diagnosis using DNA microarray data.

Availability: Software is available upon request.

Contact: honda@nubio.nagoya-ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

To propose an optimal and individualized treatment protocol of malignant diseases, such as cancer, marker gene selection is essential in the analysis of gene expression data. Employing filtering methods is necessary for marker selection. In this approach, prior to the application of mining algorithms, genes are selected by filtering methods such as Mann–Whitney's *U*-test, Student's *t*-test (Sttest), Welch's *t*-test (Wttest), signal-to-noise (S2N) (Golub *et al*., 1999), significance analysis of microarrays (SAM) (Tusher *et al*., 2001) and nearest shrunken centroids (NSC) (Tibshirani *et al*., 2002).

In our previous study, we developed the projective adaptive resonance theory (PART) filtering method further by modifying it (Cao and Wu, 2002, 2004) and reported that PART showed a better performance than conventional methods such as S2N and NSC (Takahashi *et al*., 2005). In the PART algorithm, the genes that have a low variance in the gene expression level in either of the two classes are selected. In the present study, we developed simple and practical filtering methods such as modified S2N (S2N′); this was derived from a conventional method, such as S2N, based on the concept underlying the PART filtering method. Several conventional methods such as S2N, Sttest, Wttest, SAM and NSC were modified. These filtering methods were applied to four gene expression profile datasets, and the reproducibility of prediction was evaluated to demonstrate that the new filtering methods (derived from the conventional methods) were statistically superior to the original conventional methods.

2 MATERIALS AND METHODS

2.1 Data processing

We used four kinds of gene expression profiles, such as those for leukemia, as reported by Golub *et al*. (1999); for a CNS tumor, as reported by Pomeroy *et al*. (2002); for lymphoma, as reported by Shipp *et al*. (2002) and for breast cancer, as reported by van't Veer *et al*. (2002). After preprocessing each dataset, we used 2456 genes for leukemia, 2695 genes for the CNS tumor, 2821 genes for lymphoma and 13 547 genes for breast cancer in the present study.

2.2 Evaluation of performance for each filtering method

The two datasets, namely, the first and second sets, were randomly prepared from an original dataset. The correlation coefficient of the two datasets was calculated as the correlation between the two sets of *U*-test *p*-values of each

*To whom correspondence should be addressed.

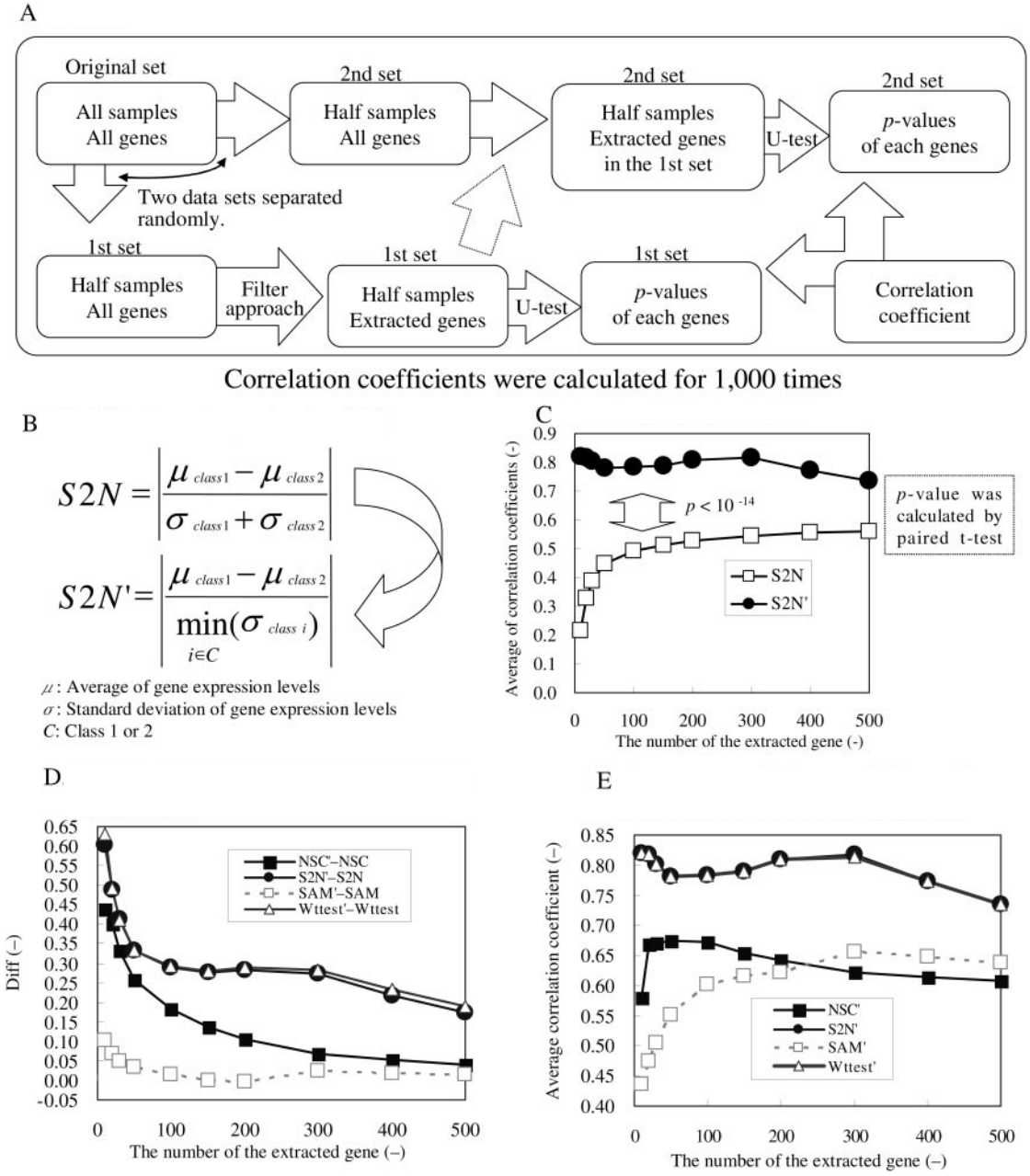

Fig. 1. Modification based on the concept of the PART filtering method. (**A**) Outline. (**B**) Modification of equation for S2N. (**C**) Results of S2N and S2N′. (**D**) Effects of modification of the equations for leukemia. (**E**) Average correlation coefficients of the medical method for leukemia.

dataset, as shown in Figure 1A. This procedure was repeated 1000 times. The S2N′ was derived from S2N, as shown in Figure 1B. For two-class discrimination, the genes that have a low variance in gene expression levels in either class, not both classes, can be selected using this modification. The performances of each method, including S2N and S2N′, were evaluated using average correlation coefficients, as shown in Figure 1C. The top 10, 20, 30, 50, 100, 150, 200, 300, 400 and 500 genes were extracted from the ranked genes by each filtering method in order to estimate differences in the correlation coefficients between the modified filtering methods and the original methods by using *p*-values of the paired *t*-test. Other conventional methods such as Sttest, Wttest, SAM and NSC were also modified.

3 RESULTS AND DISCUSSION

3.1 Comparison of the performance of the modified filter methods and original conventional methods

In order to evaluate the effects of the modification, correlations between the sets of *U*-test *p*-values in randomly divided datasets were calculated for each method and for four data sets, namely, those for leukemia, a CNS tumor, lymphoma, for breast cancer. The results of each filtering method for leukemia, are shown in Figure 1D. The results for Sttest and Sttest′ are not shown because

these are almost the same as the results for Wttest and Wttest′. Improvement in the performance after modification was evaluated using the Diff index. In this index, the average correlation coefficient of the original method is subtracted from that of the modified method. The Diff values were confirmed using paired *t*-test, and the *p*-values of all methods were $<10^{-14}$. Among various methods, the average correlation coefficient for S2N and Wttest in particular, greatly improved because of the modification. In the case of the leukemia data and top 10 genes, the diff values for S2N′ − S2N and Wttest′ − Wttest attained a value >0.6; this value corresponds to ∼300% of the average correlation coefficients for S2N and Wttest. The modifications were effective for almost all methods and data in the present study, as shown in Figure S1 (see Supplementary information). In the present study, the equations of several conventional methods such as S2N, Sttest, Wttest, SAM and NSC were modified to select the genes that have a low variance in the gene expression level in either of the two classes. Correlation coefficients of *p*-values of all methods had improved. These results showed these modified equations were more practical than conventional ones for gene expression data.

The average correlation coefficients for the modified filtering methods for leukemia are shown in Figure 1E. The average correlation coefficients for S2N′ and Wttest′ were higher than the other methods. It was concluded that S2N′ and Wttest′ are the optimal filtering methods in the present study because their equations are almost identical. Similar results were obtained for other datasets, as shown in Figure S2 (see Supplementary information).

3.2 Classification for heterogeneous samples

Generally, gene expression data obtained from cancer patients show considerable heterogeneity. However, conventional filtering methods have been designed under the hypothesis that each class has homogeneous samples; this is for the extraction of the genes that have low variances and a high difference, as shown in Figure S3A (see Supplementary information). The number of such genes in the case of heterogeneous samples is less. Therefore, many conventional methods are not practical for analyzing gene expression data of cancer patients.

The filtering methods that use clustering, such as the PART filtering method, were suitable for these data because these methods could separate the homogeneous samples from the heterogeneous samples. The three cases wherein a class has heterogeneous samples are shown in Figure S3B–D. These methods could extract the genes for all three cases appropriately. However, these methods are heuristic and time consuming. According to the results of the PART filtering method reported by Takahashi *et al*. (2005), the genes that have low variance in the gene expression level in either class are important. That is, there are many genes corresponding to two cases—Cases 2 and 3—among the three cases. Therefore, we have modified many conventional methods to extract these genes. Finally, this modification was observed to be very effective in the present study.

4 CONCLUSIONS

In the present study, we developed new filtering methods, namely, S2N′, Wttest′, Sttest′, SAM′ and NSC′, by modifying the conventional methods S2N, Wttest, Sttest, SAM and NSC, respectively, for analysis of gene expression data obtained from cancer patients. This modification is effective for almost all conventional methods. In particular, S2N′, Wttest′ and Sttest′ showed high performance owing to the modification. These results suggest that the our modification is useful and S2N′, Wttest′ and Sttest′ have high potential as new methods for marker gene selection in cancer diagnosis by using high-dimensional data such as those obtained from DNA microarrays, mass spectrometry (MS) and two-dimensional polyacrylamide gel electrophoresis.

ACKNOWLEDGEMENTS

This research was supported in part by Grant in-Aid for Scientific Research from the Japan Society for the Promotion of Science (No. 17206082). Funding to pay the Open Access publication charges for this article was provided by Ministry of Education, Science, Sports and Culture, Grant in aid for JSPS Fellows, 18·6550, 2006.

Conflict of Interest: none declared.

REFERENCES

Cao,Y. and Wu,J. (2002) Projective ART for clustering data sets in high dimensional spaces. *Neural Netw*., **15**, 105–120.

Cao,Y. and Wu,J. (2004) Dynamics of projective adaptive resonance theory model: the foundation of PART algorithm. *IEEE Trans. Neural Netw*., **15**, 245–260.

Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Pomeroy,S.L. *et al*. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Shipp,M.A. *et al*. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med*., **8**, 68–74.

Takahashi,H. *et al*. (2005) Construction of robust prognostic predictors by using projective adaptive resonance theory as a gene filtering method. *Bioinformatics*, **21**, 179–186.

Tibshirani,R. *et al*. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Tusher,V.G. *et al*. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

van't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.