JOURNAL OF BIOSCIENCE AND BIOENGINEERING
Vol. 105, No. 1, 81–84. 2008
DOI: 10.1263/jbb.105.81

Proposal of New Gene Filtering Method, BagPART, for Gene Expression Analysis with Small Sample

Takashi Kawamura,[1] Hiro Takahashi,[1§] and Hiroyuki Honda[1,2]*

Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan[1] and MEXT Innovative Research Center for Preventive Medical Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan[2]

Received 25 July 2007/Accepted 8 October 2007

A significant problem in gene expression analysis is that the sample size is substantially lower than the number of genes. Bagging is an effective method of solving this problem in the case of small sample datasets. We have devised a combination method, called the BagPART filtering method, that uses the projective adaptive resonance theory (PART) to select important genes and achieve a binary classification more accurately ($p<10^{-10}$) than conventional methods, particularly when the sample size is small.

[**Key words:** gene expression analysis, bootstrap aggregating, projective adaptive resonance theory, gene filtering method, boosted fuzzy classifier with SWEEP]

In gene expression analysis, filtering methods are needed to avoid the problem of dimensions. A number of methods have been developed to solve this problem; these include signal-to-noise (S2N) measurement, significance analysis of microarrays (SAM), and nearest shrunken centroid (NSC) (1–3). We developed a projective adaptive resonance theory (PART) filtering method for gene expression analysis by modifying the original PART filtering method reported in our previous studies (4, 5). Currently, we can obtain gene expression information from more than 10,000 genes owing to advanced DNA microarray technology; however, the sample size is, at most, about 100 in most datasets. Thus, the sample size is much smaller than the number of genes. We have improved the PART filtering method by introducing the idea of bootstrap aggregating (Bagging) (6). We have designated the new method as BagPART. We applied BagPART to the analysis of the two gene expression profile datasets and compared its results with those of PART. In addition, we tested the performance of both methods in the case of a small sample size. As a result, we have shown that BagPART is statistically superior to PART for each dataset, and that BagPART is more effective when sample size is small.

We used two sets of gene expression profiles downloaded from the Stanford Microarray Database (http://genome-www5.stanford.edu). The first set consisted of the colon cancer gene expression profiles reported by Alon *et al.* (7). This dataset comprised 2000 genes and 62 samples (40 tumor samples and 22 normal samples). The second set consisted of the prostate cancer gene expression profiles reported by Singh *et al.* (8). This dataset comprised 12,600 genes and 102 samples (52 tumor samples and 50 normal samples). We selected only those genes for which all the 102 samples showed a positive intensity, resulting in the selection of 1820 genes. Signal values were transformed to a common logarithm. For the colon dataset, because the tumor sample number is about twice as large as that of normal controls, we randomly selected tumor samples to equal the number of normal samples.

BagPART modifies the PART algorithm in the following manner. Assume that we have an original dataset $S=\{(x_1, y_1), ..., (x_n, y_n)\}$ of size n, where x_i represents the feature and y_i is the class label of the observation (x_i, y_i). We draw bootstrap datasets with the replacement $S_b^*=\{(x_1^*, y_1^*), ..., (x_n^*, y_n^*)\}$, $b=1, ..., B$ for a large B (between 50 and 200). In the bootstrap dataset, a sample is randomly selected and some degree of overlap within the sample is permitted. For each bootstrap dataset S_b^*, we carry out PART and extract the gene group G_b. We repeat this procedure for all bootstrap datasets, S_1^*, ..., S_B^*, and finally select genes that are included in more than half of the B gene groups, G_1, ..., G_B. A schematic diagram of this process is shown in Fig. 1.

A parallel comparison of PART and BagPART was carried out. Firstly, all the samples were randomly divided into two groups: one was designated a training group, which helped to construct classification models, and the other was designated a test group, which was used to evaluate the model constructed by the training group. The distributions of normal and tumor samples were equal for both groups. To investigate the performance of each model in the cases where the number of samples in the training group was small, the training sample size was changed to three patterns. The ratios of the training sample size to the total number of samples, hereafter expressed as training sample ratio (TSR), were 0.8, 0.6, and 0.4. Table 1 shows the number of samples used in the analysis of each expression dataset. Secondly, to se-

* Corresponding author. e-mail: honda@nubio.nagoya-u.ac.jp
phone: +81-(0)52-789-3215 fax: +81-(0)52-789-3214
§ Present address: College of Bioscience and Biotechnology, Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi 487-8501, Japan.

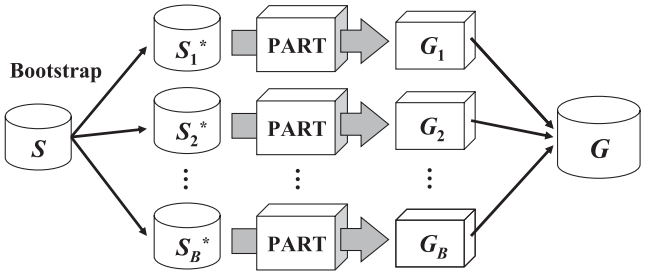

FIG. 1. Schematic image of BagPART. In the first step, the resampling data S_1*, S_2*, ..., S_B* are made using bootstrap from the original data *S*. In the second step, the genes sets G_1, G_2, ..., G_B are extracted by PART from each resampling dataset. Finally, genes included in more than half of the *B* gene sets are selected. The selected gene set is represented as *G*.

TABLE 1. Number of samples in each dataset

TSR[a]	Training sample	Test sample
Dataset: Colon cancer (normal: 22, tumor: 22)		
0.8	36	8
0.6	26	18
0.4	18	26
Dataset: Prostate cancer (normal: 50, tumor: 52)		
0.8	82	20
0.6	61	41
0.4	41	61

[a] TSR, Training sample ratio, indicates the ratio of the number of training samples to the total number of samples.

lect genes used to construct classification models, BagPART and the PART filtering method were each applied to the training group, and about 100 genes were returned by each method in both expression datasets. In the BagPART algorithm, we used $B=100$, which was adequate to converge the selected gene size (Fig. 2). Finally, classification models were constructed from the training group using only information on extracted genes and the estimated test group. Here, we used the Boosted Fuzzy Classifier with SWEEP operator method (BFCS), we developed in a previous study, which constructs 10 models for various gene combinations (9). We repeated these operations 10 times, and tested whether there is a statistically significant difference between the results using a paired t-test.

Tables 2 and 3 show the results of our analysis of the colon cancer and prostate cancer gene sets, respectively. Both results showed that BagPART was statistically superior to

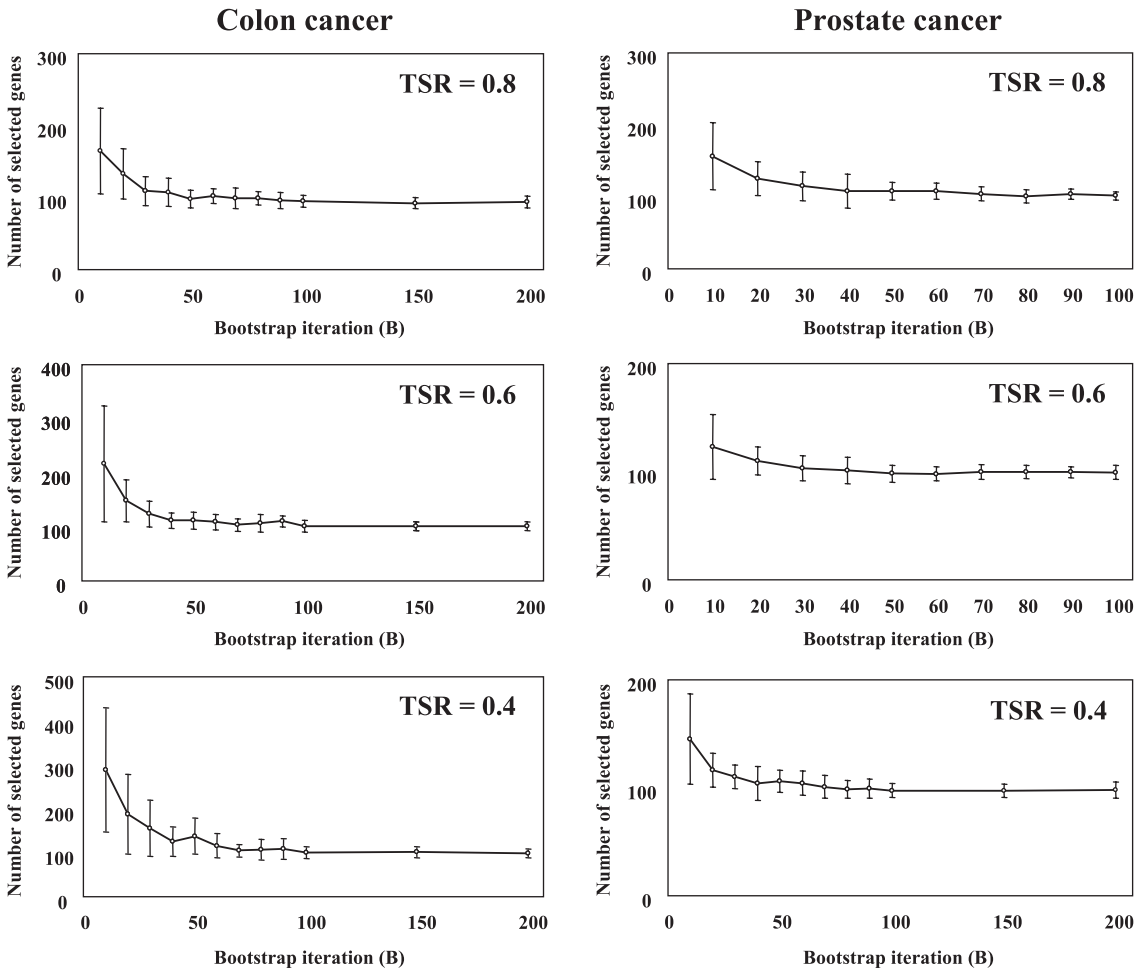

FIG. 2. Number of selected genes for each bootstrap iteration. BagPART was applied to each dataset 100 times and showed the average number of selected genes and standard deviation for each bootstrap iterations (B) in Fig. 1. The number of selected genes converged, and the standard deviation was very small in each dataset when B=100.

TABLE 2. Classification results of BagPART and PART for colon cancer dataset

TSR	Method	Average test accuracy (%)					Selected gene
		1 input	2 inputs	3 inputs	4 inputs	5 inputs	
0.8	BagPART	66.3[a]±6.6[b]	73.3±8.7	75.5±6.9	71.8±8.3	71.3±10.6	100[d]
	PART	60.3±7.2	61.0±8.3	68.1±10.0	69.0±8.3	70.9±11.2	99
	p-value			4.19×10^{-10} [c]			
0.6	BagPART	65.3±5.9	75.3±8.1	73.5±8.3	76.0±8.2	74.3±9.3	100
	PART	60.3±3.7	69.4±10.3	71.1±8.7	72.4±8.2	69.7±9.0	99
	p-value			5.30×10^{-11}			
0.4	BagPART	65.4±1.8	72.2±6.8	73.8±6.3	74.8±6.2	76.2±6.1	100
	PART	62.7±2.9	68.2±8.8	70.1±6.5	68.9±8.2	71.2±6.8	99
	p-value			2.16×10^{-14}			

[a] Average of test accuracies in 10 various datasets × 10 models.
[b] Standard deviation of test accuracies in 10 various datasets × 10 models.
[c] Result of paired t-test for comparison of test accuracies of BagPART with PART.
[d] The number of selected genes obtained from BagPART or PART. In BagPART, this is represented as *G* in Fig. 1.

TABLE 3. Classification results of BagPART and PART for prostate cancer dataset

TSR	Method	Average test accuracy (%)						Selected gene
		1 input	2 inputs	3 inputs	4 inputs	5 inputs	6 inputs	
0.8	BagPART	73.4±7.9	79.1±6.5	81.2±7.9	81.7±6.7	83.0±6.8	83.0±6.8	100
	PART	68.7±5.6	79.4±5.7	78.9±8.3	80.6±7.1	80.9±7.4	80.9±7.4	100
	p-value			1.70×10^{-5}				
0.6	BagPART	76.7±4.3	80.2±3.9	82.7±3.5	83.7±3.3	84.5±3.2	85.4±3.4	101
	PART	73.2±4.5	81.4±4.2	82.6±3.6	84.2±4.2	84.2±4.2	84.8±3.2	100
	p-value			4.58×10^{-2}				
0.4	BagPART	77.1±1.5	81.0±2.8	83.6±2.0	83.5±2.7	84.9±2.2	85.3±2.5	99
	PART	71.6±4.3	80.6±3.9	80.9±2.9	81.8±3.2	82.0±3.6	83.1±3.4	101
	p-value			2.48×10^{-13}				

PART. For the colon cancer dataset, it was clear that despite the smaller sample sizes in the training data, we observed smaller *p* values using BagPART than using PART. Thus, BagPART can correctly extract genes even if the sample size is small. For the prostate cancer dataset, when TSR decreased from 0.8 to 0.6, *p* increased $p=$from 1.70×10^{-5} to 4.58×10^{-2}. However, when TSR decreased to 0.4, *p* decreased markedly to 2.48×10^{-13}. This means that when TSR was between 0.8 and 0.6, the number of training samples in each dataset (82 samples and 61 samples, respectively), was sufficient to correctly extract genes with PART. In practice, we obtained a more significant value, $p=5.45\times10^{-15}$, when TSR decreased to 0.2 and the training sample size was 20 (data not shown). These results were in accord with previous results of Fu *et al.* (10). When the sample size is small, the performance of estimation models varies markedly depending on the proposal method. Many researchers have also reported that analytical methods using bootstrap could potentially provide more accurate estimates from datasets with small sample size.

We concluded that our method benefited from the bootstrap method. However, it is likely that when the sample size is too small, the performance of BagPART becomes poor. It is difficult to describe the minimum sample size needed for satisfactory results because the minimum sample size strongly depends on the quality of the dataset or complexity of the acquired model.

A number of genes appeared frequently in the outputs of the BagPART analysis of both datasets (Table 4). For the colon cancer dataset, in particular, the two genes, *MYH* and *COL11A2*, were frequently selected. *COL11A2* is one of two genes that encode two alpha strands of type11 collagen. *COL11A1* has been reported to be not expressed in a normal colon but to be upregulated in colorectal cancer (11). This gene was selected frequently in BagPART analysis when the TSRs were 0.8, 0.6, and 0.4. However, in regular PART analysis, this gene was only selected when TSR was 0.4, and was only the 11th most frequent gene to appear in the output (data not shown). *MYH* is relevant to DNA repair; *MYH* mutation causes familial adenomatous polyposis (FAP) (12, 13). This gene was also constantly selected in BagPART analysis regardless of TSR; however, in PART, it was not selected at all when TSR was 0.8.

MYH and *COL11A2* have been reported as important genes by many researchers, such as Le *et al.* (14), Shevade and Keerthi (15), Chu *et al.* (16), and Ma and Huang (17). However, when we employed PART as a filtering method, both genes were not frequently selected amongst all datasets. Thus, there is a substantial difference in results between PART and BagPART. For the prostate cancer dataset, *HPN* was frequently selected by both filtering methods. This gene has been reported as a marker of prostate cancer (20). Moreover, other genes, such as *AKR1B1* and *HSPD1*, which have been reported to play a role in prostate cancer (21), were frequently selected by BagPART, but not by PART (data not shown). These results indicate that BagPART can select meaningful genes even if the sample size is small. Moreover, BagPART could constantly select important genes in spite of the poor TSR values; we speculate that this characteristic is a result of using the bootstrap method.

TABLE 4. Gene lists selected frequently by BagPART

Rank	Gene ID	Genbank	Gene name	Description	Reference
Colon cancer					
1	Hsa.37937	R87126	MYH	Myosin heavy chain, nonmuscle (*Gallus gallus*)	16, 17
2	Hsa.6814	H08393	COL11A2	Collagen alpha 2(XI) chain (*Homo sapiens*)	16, 17
3	Hsa.549	R36977	GTF3A	general transcription factor Ⅲ A	18, 19
4	Hsa.627	M26383	IL8	MONAP mRNA, complete cds.	17, 18
5	Hsa.21562	R08021	PPA1	Inorganic pyrophosphatase (*Bos taurus*)	–
Prostate cancer					
1	37639_at	X07732	HPN	Hepsin (transmembrane protease, serine 1)	19, 22
2	38406_f_at	AI207842	–	Homo sapiens cDNA, 3 end	17, 19
3	36589_at	X15414	AKR1B1	Aldo-keto reductase family 1, member B1 (aldose reductase)	19
4	40282_s_at	M84526	DF	D component of complement (adipsin)	19, 22
5	37720_at	M22382	HSPD1	Heat shock 60kDa protein 1 (chaperonin)	19, 23

We have developed an improved PART filtering method that uses Bagging. To investigate the effect of the new method, BagPART, on the gene expression analysis, we applied BagPART to two types of dataset, and obtained a significant difference between the two datasets. In addition, we have clarified that the new method could more correctly extract genes than conventional methods when the sample size is small. In this study, we compared BagPART with only PART. In our previous paper, we have reported that PART is superior to signal-to-noise (S2N) measurement and nearest shrunken centroid (NSC) (4). Therefore, we believe that BagPART has better performance than the other methods, although further comparison with other methods such as SOM has not yet been carried out. We believe that, in the case of precious or rare samples, our method is advantageous for extracting important genes.

REFERENCES

1. **Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S.:** Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, **286**, 531–537 (1999).
2. **Tusher, V. G., Tibshirani, R., and Chu, G.:** Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. USA, **98**, 5116–5121 (2001).
3. **Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G.:** Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA, **99**, 6567–6572 (2002).
4. **Takahashi, H., Kobayashi, T., and Honda, H.:** Construction of robust prognostic predictors by using projective adaptive resonance theory as a gene filtering method. Bioinformatics, **21**, 179–186 (2005).
5. **Cao, Y. and Wu, J.:** Projective ART for clustering data sets in high dimensional spaces. Neural Netw., **15**, 105–120 (2002).
6. **Breiman, L.:** Bagging predictors. Mach. Learn., **24**, 123–140 (1996).
7. **Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J.:** Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA, **96**, 6745–6750 (1999).
8. **Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., and other 5 authors:** Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, **1**, 203–209 (2002).
9. **Takahashi, H. and Honda, H.:** A new reliable cancer diagnosis method using boosted fuzzy classifier with SWEEP operator method. J. Chem. Eng. Jpn., **38**, 763–777 (2005).
10. **Fu, W. J., Carroll, R. J., and Wang, S.:** Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics, **21**, 1979–1986 (2005).
11. **Fisher, H., Stenling, R., Rubio, C., and Lindblom, A.:** Colorectal carcinogenesis is associated with stromal expression of COL11A1 and COL5A2. Carcinogenesis, **22**, 875–878 (2001).
12. **Varesco, L.:** Familial adenomatous polyposis: genetics and epidemiology. Tech. Coloproctol., **8**, 305–308 (2004).
13. **Ponti, G., Ponz de Leon, M., Maffei, S., Pedroni, M., Losi, L., Di Gregorio, C., Gismondi, V., Scarselli, A., Benatti, P., Roncari, B., and 6 authors:** Attenuated familial adenomatous polyposis and Muir-Torre syndrome linked to compound biallelic constitutional MYH gene mutations. Clin. Genet., **68**, 442–447 (2005).
14. **Li, Y., Campbell, C., and Tipping, M.:** Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics, **18**, 1332–1339 (2002).
15. **Shavade, S. K. and Keerthi, S. S.:** A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics, **19**, 2246–2253 (2003).
16. **Chu, W., Ghahramani, Z., Falciani, F., and Wild, D. L.:** Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics, **21**, 3385–3393 (2005).
17. **Ma, S. and Huang, J.:** Regularized ROC method for disease classification and biomarker selection with microarray data. Bioinformatics, **21**, 4356–4362 (2005).
18. **Chen, J. J., Tsai, C., Tzeng, S., and Chen, C.:** Gene selection with multiple ordering criteria. BMC Bioinformatics, **8**, 74–90 (2007).
19. **Yap, Y., Zhang, X., Ling, M., Wang, X., Wong, Y., and Danchin, A.:** Classification between normal and tumor tissues based on the pair-wise gene expression ratio. BMC Cancer, **4**, 72–88 (2004).
20. **Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., and Winslow, R. L.:** Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics, **21**, 3905–3911 (2005).
21. **Johansson, B., Pourian, M. R., Chuan, Y. C., Byman, I., Berqh, A., Panq, S. T., Norstedt, G., Berqman, T., and Pousette, A.:** Proteomic comparison of prostate cancer cell lines LNCaP-FGC and LNCaP-r reveals heatshock protein 60 as a marker for prostate malignancy. Prostate, **66**, 1235–1244 (2006).
22. **Niijima, S. and Kuhara, S.:** Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE. BMC Bioinformatics, **7**, 543–560 (2006).
23. **Lai, Y., Wu, B., Chen, L., and Zhao, H.:** A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics, **20**, 3146–3155 (2004).