JOURNAL OF BIOSCIENCE AND BIOENGINEERING
Vol. 102, No. 1, 46–52. 2006
DOI: 10.1263/jbb.102.46

Classification of Intramural Metastases and Lymph Node Metastases of Esophageal Cancer from Gene Expression Based on Boosting and Projective Adaptive Resonance Theory

Hiro Takahashi,[1,2,3] Kazuhiko Aoyagi,[3] Yukihiro Nakanishi,[4] Hiroki Sasaki,[3]
Teruhiko Yoshida,[3] and Hiroyuki Honda[2]*

Research Fellow of the Japan Society for the Promotion of Science (JSPS), 8 Ichibancho, Chiyoda-ku, Tokyo 102-8472, Japan,[1] Department of Biotechnology, School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan,[2] Genetics Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan,[3] and Pathology Division, National Cancer Center Research Institute, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan[4]

Received 22 February 2006/Accepted 15 April 2006

Esophageal cancer is a well-known cancer with poorer prognosis than other cancers. An optimal and individualized treatment protocol based on accurate diagnosis is urgently needed to improve the treatment of cancer patients. For this purpose, it is important to develop a sophisticated algorithm that can manage a large amount of data, such as gene expression data from DNA microarrays, for optimal and individualized diagnosis. Marker gene selection is essential in the analysis of gene expression data. We have already developed a combination method of the use of the projective adaptive resonance theory and that of a boosted fuzzy classifier with the SWEEP operator denoted PART-BFCS. This method is superior to other methods, and has four features, namely fast calculation, accurate prediction, reliable prediction, and rule extraction. In this study, we applied this method to analyze microarray data obtained from esophageal cancer patients. A combination method of PART-BFCS and the U-test was also investigated. It was necessary to use a specific type of BFCS, namely, BFCS-1,2, because the esophageal cancer data were very complexity. PART-BFCS and PART-BFCS with the U-test models showed higher performances than two conventional methods, namely, *k*-nearest neighbor (kNN) and weighted voting (WV). The genes including *CDK6* could be found by our methods and excellent IF-THEN rules could be extracted. The genes selected in this study have a high potential as new diagnosis markers for esophageal cancer. These results indicate that the new methods can be used in marker gene selection for the diagnosis of cancer patients.

[**Key words:** cancer classification, boosting, projective adaptive resonance theory, esophageal cancer, intramural metastases]

Cancer is a major cause of human deaths in the many countries. Esophageal cancer is the eighth most common cancer and the sixth most common cause of cancer-related mortality in the world (1). This cancer is a well-known cancer with poorer prognosis than other cancers. Lymph node metastasis is one of the reasons for its poor prognosis in potentially resectable solid epithelial tumors. Furthermore, intramural metastasis (skip metastasis) has poorer prognosis than lymph node metastasis (2). From such situations, the prognosis of cancer patients with the same clinical diagnosis can differ, frequently. Therefore, it is important that the prognosis of cancer patients is made accurately and that an adequate treatment is proposed. However, the diagnosis of cancer patients is determined by a complex causality involving multiple factors because the mechanisms of cancer development (or malignancy) are extremely complex. Gene expression data from DNA microarrays are individualized and useful in the diagnosis and prognosis of diseases (3). To conduct this analysis, it is necessary to select genes significantly expressing mRNA and strongly related to the diagnosis or prognosis of disease, because the performance of classification analysis can decline owing to such large quantities of data.

Feature selection has been performed to screen candidate genes for modeling. There are two types of approach: the wrapper and filter approaches. In the former, features (genes) are selected as a part of mining algorithms, such as support vector machines (SVMs) (4), a fuzzy neural networks (FNNs) combined with the SWEEP operator method (FNN-SWEEP) (3), and a boosted fuzzy classifier with the SWEEP operator method (BFCS) (5, 6). On the other hand, in the filter approach, features are selected by filtering methods, such as the U-test, the t-test, signal-to-noise statistic (S2N)

* Corresponding author. e-mail: honda@nubio.nagoya-u.ac.jp
phone: +81-(0)52-789-3215 fax: +81-(0)52-789-3214

(7) and the projective adaptive resonance theory (PART) (8), prior to the application of mining algorithms.

In our previous study, we investigated the combinations of various filter and wrapper approaches and applied these combination methods to microarray data of acute leukemia and central nervous system tumors (CNS). Consequently, we showed that a combination method of the use of projective adaptive resonance theory and that of a boosted fuzzy classifier with the SWEEP operator method denoted PART-BFCS was the best among various combination methods for constructing on accurate model resulting in an accurate prediction. In this study, we applied this method to the analysis of expression profile data of esophageal cancer. In addition, the performances of BFCS or PART-BFCS with the U-test models, were investigated. The constructed PART-BFCS with the U-test or PART-BFCS models could accurately discriminate esophageal cancer patients with intramural metastases (IMs) from other esophageal cancer patients, BFCS with the U-test (U-test-BFCS) models could not.

It is necessary to select specific and essential marker genes for cancer classification and diagnosis. Minimum gene sets without false positive ones should be extracted. Therefore, various methods were compared under the condition of small inputs. We concluded that our method is the best under this condition for esophageal cancer analysis.

MATERIALS AND METHODS

Microarray analysis Gene expression profile data were obtained from 64 surgical specimens from esophageal cancer patients: 16 patients who had no lymph node metastases (O1), 6 patients who had lymph node metastases from one to four (O2), 29 patients who had over four lymph node metastases (O3), and 13 patients who had some IMs (see Table 1A). For RNA extraction, trained pathologists carefully excised bulk tissue samples from the main tumor, leaving a clear margin from the surrounding nontumorous tissue. Total RNAs extracted from the bulk tissue samples were biotin-labeled and hybridized to high-density oligonucleotide microarrays (Affymetrix Human Genome U95A Array) containing 12,600 probe sets representing 10,000 transcripts according to the manufacturer's instructions. The scanned data of the arrays were processed by Affymetrix Microarray Suite, which scaled the average intensity of all the genes on each array to a target signal of 1000.

Data processing As shown in Table 1B, the esophageal cancer data were partitioned into two data sets: 54 samples (42 non-IM and 12 IM) as a modeling data set for constructing the class prediction model (predictor) and 10 samples (9 non-IM and 1 IM) as a blind data set for evaluating the constructed predictor (10 blind data), and a leave-one-out cross-validation set (LOOCV data). We excluded genes expressed at a P call (meaning expression signal is present) of less than 10 in the 64 specimens. As a result, 8037 probes were selected in this preprocessing step. During the gene- filtering step, 1000 probes were selected using PART and the U-test, respectively, and then two types of BFCS, namely, BFCS-1 and BFCS-1,2 were used in the modeling step as wrapper approaches. For comparison, conventional modeling methods without filtering, namely, weighted voting (WV) (7) and *k*-nearest neighbor (kNN), were also used.

kNN method The *k*-nearest neighbor (kNN) method is based on a distance function for pairs of tumor samples, such as Euclidean distance. kNN proceeded as follows to classify blind data set observations on the basis of the modeling data set. For each patient in the blind data set (i) the *k* closest neighbors in the modeling data set were found, and (ii) class was predicted by majority vote; that is, the class that is most common among those *k* neighbors was chosen. The number of neighbors ($k=3$) was used because a similar cross-validation accuracy of models was obtained in the modeling data set for various *k*s.

WV method WV was originally proposed by Golub *et al.* (7) to manage microarray data. The weight of each gene was calculated using signal-to-noise statistic. The linear models of one gene were assembled by gene weight.

Model construction with parameter selection The parameter increasing method (PIM) (9) was used to select input combinations for the construction of kNN and WV models. This was performed as follows.

First, we predicted the class (IM or non-IM) of each sample using the prediction model with a single input. Prediction models for each probe were constructed in series, and all the probes were ordered on the basis of the accuracy of the constructed models. In the next step, the probe with the highest accuracy was used to construct a combination model.

Second, we selected a partner probe for the probe selected in the first step to increase prediction accuracy. To accomplish this, we

TABLE 1. List of esophageal cancer patients

A. All patients

Stage of metastasis	Description	Number of patients
O1	Lymph node metastases = 0	16
O2	4 ≥ Lymph node metastases ≥ 1	6
O3	Lymph node metastases > 4	29
IM	Intramural metastases (IM)	13
	Total	64

B. Divided data set

Data set name	Stage of metastasis	Content of data blocks		Number of data blocks
		Number in the modeling data	Number in the blind data	
Blind 10 data	Non-IM (O1, O2, O3)	42	9	1
	IM	12	1	
Leave-one-out	Non-IM	51	0	13
Cross-validation	IM	12	1	
(LOOCV) data	Non-IM	50	1	51
	IM	13	0	

constructed a 2-input model in which a ranked probe was designated input 1, and input 2 (partner probe) was selected to provide the highest training accuracy while applying kNN (or WV) and PIM to the analysis of the modeling data. By repeating this step, an optimum combination of $N_{\mathrm{attribute}}$ candidate probes was identified for use as input probes in the model construction. $N_{\mathrm{attribute}}$ was defined as ten in this study.

Finally, combinations of $N_{\mathrm{attribute}}$ probes, *i.e.*, from the first to the $N_{\mathrm{attribute}}{}^{\mathrm{th}}$ probes were evaluated. We constructed $N_{\mathrm{attribute}}$ predictor models, beginning with one input using only the first-selected probe to $N_{\mathrm{attribute}}$ inputs using all the $N_{\mathrm{attribute}}$ probes. The performance of the prediction models was evaluated by applying them to the analysis of the blind data set.

For the two data sets, the genes with the 1st to the 10th highest accuracies were used as the first inputs for the construction of the 10 combination models by PIM.

BFCS method Boosting was proposed by Schapire (10), and thus far, several derivative boosting algorithms (11–13) have been developed. Boosting is useful for class prediction using high-dimensional inputs and very fast algorithms.

In our previous study, we developed a boosted fuzzy classifier with the SWEEP operator method (BFCS) (5) on the basis of AdaBoost (11), which is the most basic boosting algorithm. This method enables the evaluation of the reliability of the predictions for each patient. However, it is difficult to evaluate the reliability of the predicted results of conventional boosting.

A BFCS model is composed of type I fuzzy neural network (FNN) models (14). In this study, 1- or 2-input FNN models were used as weak learners in the BFCS model, and they were combined with connection weights, which were determined using the AdaBoost algorithm. BFCS has two types, BFCS-1 and BFCS-1,2. A BFCS-1 model is composed of 1-input FNN models (5). On the other hand, BFCS-1,2 is composed of 1- or 2-input FNN models (5). BFCS-1,2 can used for analyzing the interaction between two inputs, because this method can includes 2-input FNN models.

PART-BFCS Previously, we developed and combined the use of the projective adaptive resonance theory (PART) as a gene filtering method and that of a boosted fuzzy classifier with the SWEEP operator method (BFCS) as a modeling method. In the resulting method PART-BFCS, PART first preselects the genes that show small variances within a class. Then, BFCS rapidly selects these genes to build a highly accurate and reliable predictor.

PART has two important parameters, vigilance and distance. Vigilance was optimized so that modeling samples clustered well. Distance was used to control the number of extracted genes. The genes extracted by PART showed a low standard deviation (SD) in the low-gene–expression-level class. The predictor using genes with a low SD in low class showed a high performance (8).

In BFCS, 1- or 2-input FNN models based on the neural network and fuzzy logic were used as weak learners. The BFCS models constructed using only 1-input FNN models were defined as a BFCS-1 model, and those constructed using 1- or 2-input FNN models were defined as a BFCS-1,2 model in our previous study.

RESULTS AND DISCUSSION

Selection of BFCS type and complexity of esophageal cancer data for the classification of IM and non-IM BFCS-1 is effective for analyzing many gene expression profiles, such as those of acute leukemia, central nervous system tumors (CNS), and soft tissue sarcomas (unpublished data). BFCS-1 without screening was applied to the analysis of the modeling data of esophageal cancer shown in Fig. 1. Figure 1 shows training curves against the number of genes. The solid line indicates the training curve for the esophageal cancer data. The dashed lines indicate the modeling results for other cancer data, namely, the acute leukemia, and CNS data. The training curve result obtained by the BFCS-1 expressed underfitting of the esophageal cancer data, and a training curve result of 100% was achieved for the data of the other two cancers. This result implies that the esophageal cancer data were very complex. Therefore, BFCS-1,2 was used in this study, because it is more effective than BFCS-1 in the cases in which the relationships of the attributes provided and its output are highly complex.

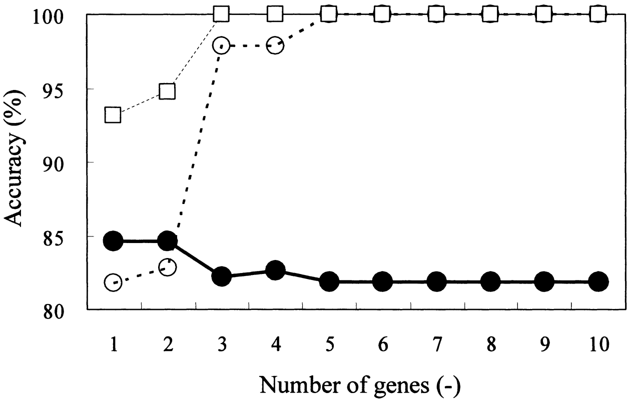


FIG. 1. Training curves of BFCS-1 without screening for modeling data of 10 blind data. The training curves were developed using average training accuracy from 10 combination models constructed by BFCS-1. The solid line with filled circles is the training curve for the esophageal cancer data. The dashed line with open circles is the curve for the acute leukemia data. The dashed line with open squares is the curve for the central nervous system (CNS) tumor data. The leukemia and CNS data were obtained from the website http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

Comparison of performances of BFCS with filtering methods with those of other methods The performances of BFCSs with filtering methods as models were investigated, namely, BFCS with PART (PART-BFCS), BFCS with the U-test (U-test-BFCS), and BFCS with PART and the U-test (PART-BFCS with U-test). For comparison, the predictors of two conventional methods, namely, WV and kNN, were constructed. The performances of the predictors were compared in terms of accuracy using a blind data set that was not used for modeling. By using 10 combination models, the average accuracy for the blind data set was calculated for the two data sets, namely, 10 blind and LOOCV data.

Results of LOOCV data are shown in Table 2. The results show that the average accuracy of 6-input PART-BFCS with the U-test models is the highest. The average accuracies of the BFCSs with filtering methods were higher than those of two conventional methods, namely, WV and kNN. However, U-test-BFCS models showed a very low sensitivity.

Results of 10 blind data are shown in Table 3. The results show that the average accuracy of 10-input PART-BFCS with the U-test methods is the highest and that the average accuracies of models for BFCS with filtering methods were higher than those of the conventional methods. However, U-test-BFCS model also shows a very low sensitivity.

A comparison of PART-BFCS and PART-BFCS with the

TABLE 2. Comparison of performances of various methods for LOOCV data

	Method (–)	Inputs (–)									
		1	2	3	4	5	6	7	8	9	10
Accuracy (%)	BFCS with PART and U-test	–	75.0	–	75.8	–	80.9[a]	–	78.8	–	80.3
	BFCS with PART	–	76.4	–	75.8	–	77.2	–	77.3	–	78.1
	BFCS with U-test	–	65.5	–	68.6	–	73.0	–	73.0	–	76.1
	kNN	74.7	70.0	70.8	70.2	71.3	69.5	70.3	68.1	69.4	69.1
	WV	61.3	64.1	66.1	69.8	63.0	62.2	63.6	65.9	65.9	64.7
Sensitivity (%)	BFCS with PART and U-test	–	15.4	–	21.5	–	21.5	–	13.1	–	11.5
	BFCS with PART	–	16.2	–	25.4	–	16.9	–	13.1	–	6.2
	BFCS with U-test	–	2.3	–	3.8	–	0.0	–	0.0	–	0.0
	kNN	23.8	24.6	25.4	20.8	21.5	18.5	16.2	14.6	19.2	16.9
	WV	14.6	12.3	13.8	16.9	15.4	19.2	17.7	16.2	15.4	16.2
Specificity (%)	BFCS with PART and U-test	–	90.2	–	89.6	–	96.1	–	95.5	–	97.8
	BFCS with PART	–	91.8	–	88.6	–	92.5	–	93.7	–	96.5
	BFCS with U-test	–	81.6	–	85.1	–	91.6	–	91.6	–	95.5
	kNN	87.6	81.6	82.4	82.7	83.9	82.5	84.1	81.8	82.2	82.4
	WV	73.1	77.3	79.4	83.3	75.1	73.1	75.3	78.6	78.8	77.1

[a] The highest accuracy. – indicates that no models were constructed, because BFCS-1,2 method selected a 2-input weak learner consisting of two genes. Accuracy is the ratio of correctly predicted patients to total patients. Sensitivity is accuracy for IM patients. Specificity is accuracy for non-IM patients.

TABLE 3. Comparison of performances of various methods for 10 blind data

	Method (–)	Inputs (–)									
		1	2	3	4	5	6	7	8	9	10
Accuracy (%)	BFCS with PART and U-test	–	80.0	–	84.0	–	85.0	–	89.0	–	96.0[a]
	BFCS with PART	–	83.0	–	81.0	–	82.0	–	83.0	–	88.0
	BFCS with U-test	–	82.0	–	79.0	–	84.0	–	83.0	–	88.0
	kNN	72.0	74.0	72.0	80.0	77.0	75.0	78.0	76.0	73.0	69.0
	WV	66.0	67.0	57.0	60.0	65.0	62.0	70.0	65.0	61.0	64.0
Sensitivity (%)	BFCS with PART and U-test	–	50.0	–	60.0	–	80.0	–	80.0	–	80.0
	BFCS with PART	–	70.0	–	80.0	–	90.0	–	90.0	–	90.0
	BFCS with U-test	–	30.0	–	10.0	–	10.0	–	0.0	–	0.0
	kNN	20.0	0.0	10.0	10.0	10.0	10.0	20.0	20.0	20.0	0.0
	WV	30.0	40.0	20.0	10.0	50.0	30.0	30.0	0.0	20.0	40.0
Specificity (%)	BFCS with PART and U-test	–	83.3	–	86.7	–	85.6	–	90.0	–	97.8
	BFCS with PART	–	84.4	–	81.1	–	81.1	–	82.2	–	87.8
	BFCS with U-test	–	87.8	–	86.7	–	92.2	–	92.2	–	97.8
	kNN	77.8	82.2	78.9	87.8	84.4	82.2	84.4	82.2	78.9	76.7
	WV	70.0	70.0	61.1	65.6	66.7	65.6	74.4	72.2	65.6	66.7

[a] The highest accuracy. – indicates that no models were constructed, because BFCS-1,2 method selected a 2-input weak learner consisting of two genes. Accuracy is the ratio of correctly predicted patients to total patients. Sensitivity is accuracy for IM patients. Specificity is accuracy for non-IM patients.

U-test was performed using the accuracies of 100 models (2 data sets × 10 combination models × 5 types of input from 2 to 10). The *P* value was 0.022 and was calculated using the paired t-test. PART-BFCS with the U-test was superior to PART-BFCS for esophageal cancer data. These results indicate that PART is necessary for BFCS, because PART eliminates genes which hinder the prediction of BFCS. In addition, PART-BFCS with the U-test was the best method for analyzing esophageal cancer data.

Comparison of selected genes by PART-BFCS and PART-BFCS with U-test The average accuracy of 6-input PART-BFCS with the U-test models was the highest, as shown in Table 2. The detailed results of ten combination 6-input PART-BFCS with the U-test models were analyzed (data not shown). Results of the PART-BFCS were also analyzed, because this method had the second highest accuracy of the 6-input models. The results showed that the accuracies of all the models used are almost the same. However, sensitivity markedly differed between the models; the sensitivities ranged from 0.0% to 46.2% for PART-BFCS with the U-test models, and from 7.7% to 38.5% for PART-BFCS models. The variance in sensitivity was large, because the number of IM patients was very small in this study. Therefore, the highest sensitivity models among ten combinations for each method were selected for the following analysis; the no. 4 model for PART-BFCS with the U-test and the no. 5 model for PART-BFCS.

Actually, 99 and 121 independent genes (probe sets) were selected and the top 10 genes that were selected most frequently are shown in Table 4A. Table 4A shows that the gene *CDK6* was selected most and the gene *SIM2* was selected 2nd most for both models. *CDK6* is a well-known cell cycle regulation gene and is an important marker for cancer diagnosis (15–17). For 10 blind data, *CDK6* was also selected frequently, as shown in Table 5.

Next, we investigated the genes selected together with

TABLE 4. List of genes selected by 6-input BFCS with screening for LOOCV data

A. The selected genes

Model	Gene name	Genbank	Description	Number of times selected
No. 4 model of BFCS with PART and U-test	CDK6	X66365	Cyclin-dependent kinase 6	45
	SIM2	U80456	Single-minded homolog 2 (*Drosophila*)	27
	MYL6	M22919	Myosin, light polypeptide 6, alkali, smooth muscle and non-muscle	19
	TRIP6	AJ001902	Thyroid hormone receptor interactor 6	19
	C19orf2	AB006572	Chromosome 19 open reading frame 2	17
	FBXO21	AB020682	F-box only protein 21	13
	KCNJ15	Y10745	Potassium inwardly-rectifying channel, subfamily J, member 15	12
	ZNF3	X07290	Zinc finger protein 3 (A8-51)	11
	POLS	AB005754	Polymerase (DNA directed) sigma	11
	NFIB	AI222594	Nuclear factor I/B	10
No. 5 model of BFCS with PART	CDK6	X66365	Cyclin-dependent kinase 6	37
	SIM2	U80456	Single-minded homolog 2 (*Drosophila*)	28
	C19orf2	AB006572	Chromosome 19 open reading frame 2	18
	TRIP6	AJ001902	Thyroid hormone receptor interactor 6	16
	POLS	AB005754	Polymerase (DNA directed) sigma	13
	ERCC1	M13194	Excision repair cross-complementing rodent repair deficiency, complementation group 1 (includes overlapping antisense sequence)	13
	FZD5	U43318	Frizzled homolog 5 (*Drosophila*)	12
	ZNF3	X07290	Zinc finger protein 3 (A8-51)	12
	NFIB	AI222594	Nuclear factor I/B	10
	TIAL1	D64015	TIA1 cytotoxic granule-associated RNA binding protein-like 1	9

B. Genes selected together with *CDK6*

Model	Gene name	Genbank	Description	Number of times selected
No. 4 model of BFCS with PART and U-test	C19orf2	AB006572	Chromosome 19 open reading frame 2	17
	MYL6	M22919	Myosin, light polypeptide 6, alkali, smooth muscle and non-muscle	9
	FZD5	U43318	Frizzled homolog 5 (*Drosophila*)	4
	FBXO21	AB020682	F-box only protein 21	3
	GPA33	U79725	Glycoprotein A33 (transmembrane)	3
	TRIP13	U96131	Thyroid hormone receptor interactor 13	2
	TCF4	M74719	Transcription factor 4	2
No. 5 model of BFCS with PART	C19orf2	AB006572	Chromosome 19 open reading frame 2	18
	FZD5	U43318	Frizzled homolog 5 (*Drosophila*)	12
	TRIP13	U96131	Thyroid hormone receptor interactor 13	2

(A) The list of these genes was sorted by the number of times selected in the LOOCV (64-fold), and the top 10 genes are shown. Independent 99 and 121 genes (probe sets) were selected for each model, respectively. Except for the names of genes described, those of other 89 genes (probe sets) involved in no. 4 model and 111 genes (probe sets) involved in no. 5 model were omitted. (B) BFCS-1,2 consisted of 2-input FNN models concluding two genes. Only the genes selected two or more times are shown. Except for the names of genes described, those of other 5 genes (probe sets) involved in each no. 4 and no. 5 model were omitted.

CDK6, as shown in Tables 4B and 5. For 10 blind data, Table 5 showed that *FZD5* and *GPA33* were frequently selected together with *CDK6* gene. For LOOCV data, Table 4B showed that *C19orf2* and *FZD5* were also selected frequently.

Comparison of accuracy of 2-input models including those for *CDK6* with those of other models The performances of 1- or 2-input BFCS models were calculated and are shown in Table 6, such as those for *CDK6+C19orf2*, *CDK6+FZD5*, *CDK6+GPA33*, *CDK6*, *C19orf2*, *FZD5*, *GPA33*, *CDK6+SIM2*, and the negative control. The negative control indicates the average performance of 2-input models selected randomly 20,000 times. Table 6 shows that the accuracies and sensitivities of 2-input models, such as those for *CDK6+C19orf2*, *CDK6+FZD5*, and *CDK6+GPA33*, are very high. On the other hand, the sensitivities of 1-input models, such as those for *CDK6*, *C19orf2*, *FZD5*, and *GPA33*, were zero percent. The irrelevant 2-input models, namely, those for *CDK6+SIM2* and the negative control, showed low sensitivities. These results show that all the patients are classified as non-IM patients by all the 1-input models used, because the 1-input models could not be constructed correctly owing to the high complexity of these data. These results show that 2-input combinations of *CDK6*, such as *CDK6+C19orf2*, *CDK6+FZD5*, and *CDK6+GPA33* are very important.

IF-THEN rules extracted from BFCS model After modeling, the IF-THEN rules for esophageal cancer with

TABLE 5. List of genes selected by BFCS with screening methods for 10 blind data

Method	Inputs (–)	Order of selection	Combination no.									
			1	2	3	4	5	6	7	8	9	10
BFCS with PART and U-test	2	1	POLS	HMGN1	SPTAN1	FBXO21	SHARP	PC4	RSU1	RSU1	SIM2	HMGN1
			BIG1	PC4	MEST	SIM2	SIM2	SIM2	G2AN	SIM2	ATP6AP2	PCSK1
	4	2	DNASE1L1	DNASE1L1	FBXO21	DNASE1L1	DNASE1L1	DNASE1L1	STARD3	DNASE1L1	DNASE1L1	RSU1
			Unknown	Unknown	TRIP6	Unknown	Unknown	Unknown	RAGE	Unknown	Unknown	G2AN
	6	3	HMGN1	SEC24A	HMGN1	HMGN1	HMGN1	SEC24A	HMGN1	HMGN1	HMGN1	DNASE1L1
			PC4	BIG1	PC4	PC4	PC4	BIG1	PC4	PC4	PC4	SLC10A3
	8	4	FBXO21	CDK6[a]	CDK6[a]	CDK6[a]	CDK6[a]	ERCC1	DNASE1L1	CDK6[a]	CDK6[a]	SEC24A
			TRIP6	C19orf2	LRP5	GPA33[b]	GPA33[b]	OXCT	Unknown	GPA33[b]	GPA33[b]	BIG1
	10	5	SHARP	FBXO21	OAS1	SEC24A	SEC24A	CDK6[a]	SEC24A	SEC24A	SEC24A	Unknown
			SIM2	SIM2	NFIB	BIG1	BIG1	GPA33	BIG1	BIG1	BIG1	BTAF1
BFCS with PART	2	1	POLS	HMGN1	SPTAN1	C21orf25	FBXO21	DKFZp547K	ARCN1	ZNF294	SHARP	NMU
			BIG1	PC4	MEST	SIM2	SIM2	SIM2	SIM2	SIM2	SIM2	SIM2
	4	2	SAA1	CDK6[a]	FBXO21	DNASE1L1	DNASE1L1	DNASE1L1	DNASE1L1	DNASE1L1	DNASE1L1	DNASE1L1
			SIM2	MADH4	TRIP6	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
	6	3	CDK6[a]	CCBP2	HMGN1	HMGN1	HMGN1	HMGN1	HMGN1	HMGN1	HMGN1	HMGN1
			FLJ31564	POLS	PC4	PC4	PC4	PC4	PC4	PC4	PC4	PC4
	8	4	HMGN1	Unknown	CDK6[a]	CDK6[a]	CDK6[a]	CDK6[a]	CDK6[a]	CDK6[a]	CDK6[a]	CDK6[a]
			PC4	PRSS3	FLJ31564	FZD5[b]	FZD5[b]	FZD5[b]	FZD5[b]	FZD5[b]	FZD5[b]	FZD5[b]
	10	5	FBXO21	TERF1	OAS1	SAA1	SAA1	SAA1	SAA1	SAA1	SAA1	SAA1
			MINA53	MMP9	NFIB	BIG1	BIG1	BIG1	BIG1	BIG1	BIG1	BIG1

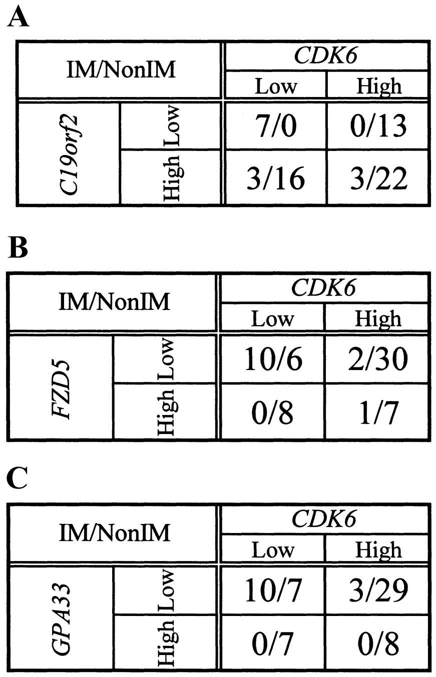
[a] *CDK6*.
[b] Genes were selected together with *CDK6*.

TABLE 6. Comparison of prediction accuracies of genes frequently selected by BFCS

Used genes (–)	Number of input	Accuracy (%)	Sensitivity (%)	Specificity (%)
CDK6+C19orf2[a]	2	89.1	53.8	98.0
CDK6+FZD5[a,b]	2	84.4	76.9	86.3
CDK6+GPA33[b]	2	82.8	76.9	84.3
CDK6	1	79.7	0.0	100.0
C19orf2	1	79.7	0.0	100.0
FZD5	1	79.7	0.0	100.0
GPA33	1	79.7	0.0	100.0
CDK6+SIM2[c]	2	79.7	30.8	92.2
Nagative control[d]	2	78.8±1.4	0.6±2.9	98.7±1.8

Accuracies were calculated by BFCS for LOOCV data.
[a] Gene that was frequently selected with *CDK6* for LOOCV data.
[b] Gene that was frequently selected with *CDK6* for 10 blind data.
[c] Gene that was the frequently selected 2nd for LOOCV data.
[d] Two genes were randomly extracted from the genes never selected by PART-BFCS or PART-BFCS with the U-test methods, and the model was constructed by BFCS. This procedure was repeated for 20,000 times.

IM and non-IM were obtained from the models including *CDK6*. The IF-THEN rules were obtained as a matrices that are classified by the expression level of selected genes for three 2-input models (Fig. 2). Using these matrices, simple and excellent rules were obtained as follows. The first rule is that patients with low expression levels of *CDK6* and *C19orf2* are likely to be IM patients, as shown in Fig. 2A. Seven patients showed low expression levels of *CDK6* and *C19orf2* and all of them were IM patients, corresponding to 54% (7/13) of all the IM patients. The next rule is that patients with low expression levels of *CDK6* and *FZD5* are likely to be IM patients, as shown in Fig. 2B. Sixteen patients showed low expression levels of *CDK6* and *FZD5* and 10 of them were IM patients, corresponding to 77% (10/13) of all the IM patients. The third rule is that patients with low expression levels of *CDK6* and *GPA33* are likely to be IM patients, as shown in Fig. 2C. Seventeen patients showed low expression levels of *CDK6* and *GPA33* and 10 of them were IM patients, corresponding to 77% (10/13) of all the IM patients. Non-IM or IM patients clustered at spe-



A

IM/NonIM		*CDK6* Low	*CDK6* High
C19orf2	Low	7/0	0/13
	High	3/16	3/22

B

IM/NonIM		*CDK6* Low	*CDK6* High
FZD5	Low	10/6	2/30
	High	0/8	1/7

C

IM/NonIM		*CDK6* Low	*CDK6* High
GPA33	Low	10/7	3/29
	High	0/7	0/8

FIG. 2. IF-THEN rules including those for *CDK6*. Because each gene can be divided into either a high or a low group using fuzzy logic, this model comprised 4 (=2^2) fuzzy rules. Values on the left in each matrix indicate the number of IM patients. Values on the right indicate the number of non-IM patients. (A) For *CDK6* and *C19orf2*. (B) For *CDK6* and *FZD5*. (C) *CDK6* and *GPA33*.

cific parts of the matrices.

In this study, we applied PART-BFCS, and PART-BFCS with the U-test to discriminate esophageal cancer patients with IM from those with non-IM. It was necessary that a specific type of BFCS, BFCS-1,2, was used, because the esophageal cancer data used were highly complex. PART-BFCS and PART-BFCS with the U-test models showed higher performances than WV and kNN. PART-BFCS with the U-test was superior to PART-BFCS. The genes including *CDK6* were found using our methods. Accurate IF-THEN rules were extracted. The genes selected in this study have a high potential as new diagnosis markers for esophageal cancer. These results indicate that these methods are new methods of marker gene selection for the diagnosis of cancer patients.

ACKNOWLEDGMENTS

We would like to thank the Hori Information Science Promotion Foundation for financial support.

This work was supported in part by the program for the promotion of Fundamental Studies in Health Sciences of the National Institute of Biomedical Innovation (NiBio) and a Grant-in-Aid for the Third Comprehensive 10-Year Strategy for Cancer Control (10 years from FY2004) and for Cancer Research (16–15) from the Ministry of Health, Labour and Welfare of Japan.

REFERENCES

1. **Parkin, D. M., Bray, F., Ferlay, J., and Pisani, P.:** Estimating the world cancer burden: Globocan 2000. Int. J. Cancer, **94**, 153–156 (2001).
2. **Igaki, H., Kato, H., Tachimori, Y., Sato, H., Daiko, H., and Nakanishi, Y.:** Prognostic evaluation for squamous cell carcinomas of the lower thoracic esophagus treated with three-field lymph node dissection. Eur. J. Cardiothorac. Surg., **19**, 887–893 (2001).
3. **Ando, T., Suguro, M., Hanai, T., Kobayashi, T., Honda, H., and Seto, M.:** Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma. Jpn. J. Cancer Res., **93**, 1207–1212 (2002).
4. **Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.:** Gene selection for cancer classification using support vector machines. Mach. Learn., **46**, 389–422 (2002).
5. **Takahashi, H. and Honda, H.:** A new reliable cancer diagnosis method using boosted fuzzy classifier with a SWEEP operator method. J. Chem. Eng. Jpn., **38**, 763–773 (2005).
6. **Takahashi, H. and Honda, H.:** Prediction of peptide binding to major histocompatibility complex class II molecules through use of boosted fuzzy classifier with SWEEP operator method. J. Biosci. Bioeng., **101**, 137–141 (2006).
7. **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).
8. **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).
9. **Noguchi, H., Hanai, T., Honda, H., Harrison, L. C., and Kobayashi, T.:** Fuzzy neural network-based prediction of the motif for MHC class II binding peptides. J. Biosci. Bioeng., **92**, 227–231 (2001).
10. **Schapire, R. E.:** The strength of weak learnability. Mach. Learn., **5**, 197–227 (1990).
11. **Freund, Y. and Schapire, R. E.:** A decision-theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci., **55**, 119–139 (1997).
12. **Friedman, J., Hastie, T., and Tibshirani, R.:** Additive logistic regression: a statistical view of boosting. Ann. Stat., **28**, 337–407 (2000).
13. **Freund, Y.:** An adaptive version of the boost by majority algorithm. Mach. Learn., **43**, 293–318 (2000).
14. **Horikawa, S., Furuhashi, T., and Uchikawa, Y.:** On fuzzy modeling using fuzzy neural networks with the back-propagation algorithm. IEEE Trans. Neural Netw., **3**, 801–806 (1992).
15. **Mendrzyk, F., Radlwimmer, B., Joos, S., Kokocinski, F., Benner, A., Stange, D. E., Neben, K., Fiegler, H., Carter, N. P., Reifenberger, G., Korshunov, A., and Lichter, P.:** Genomic and protein expression profiling identifies CDK6 as novel independent prognostic marker in medulloblastoma. J. Clin. Oncol., **23**, 8853–8862 (2005).
16. **Garcia, J. F., Camacho, F. I., Morente, M., Fraga, M., Montalban, C., Alvaro, T., Bellas, C., Castano, A., Diez, A., Flores, T., Martin, C., Martinez, M. A., Mazorra, F., Menarguez, J., Mestre, M. J., Mollejo, M., Saez, A. I., Sanchez, L., and Piris, M. A.:** Hodgkin and Reed-Sternberg cells harbor alterations in the major tumor suppressor pathways and cell-cycle checkpoints: analyses using tissue microarrays. Blood, **101**, 681–689 (2003).
17. **Henshall, S. M., Quinn, D. I., Lee, C. S., Head, D. R., Golovsky, D., Brenner, P. C., Delprado, W., Stricker, P. D., Grygiel, J. J., and Sutherland, R. L.:** Overexpression of the cell cycle inhibitor p16INK4A in high-grade prostatic intraepithelial neoplasia predicts early relapse in prostate cancer patients. Clin. Cancer Res., **7**, 544–550 (2001).