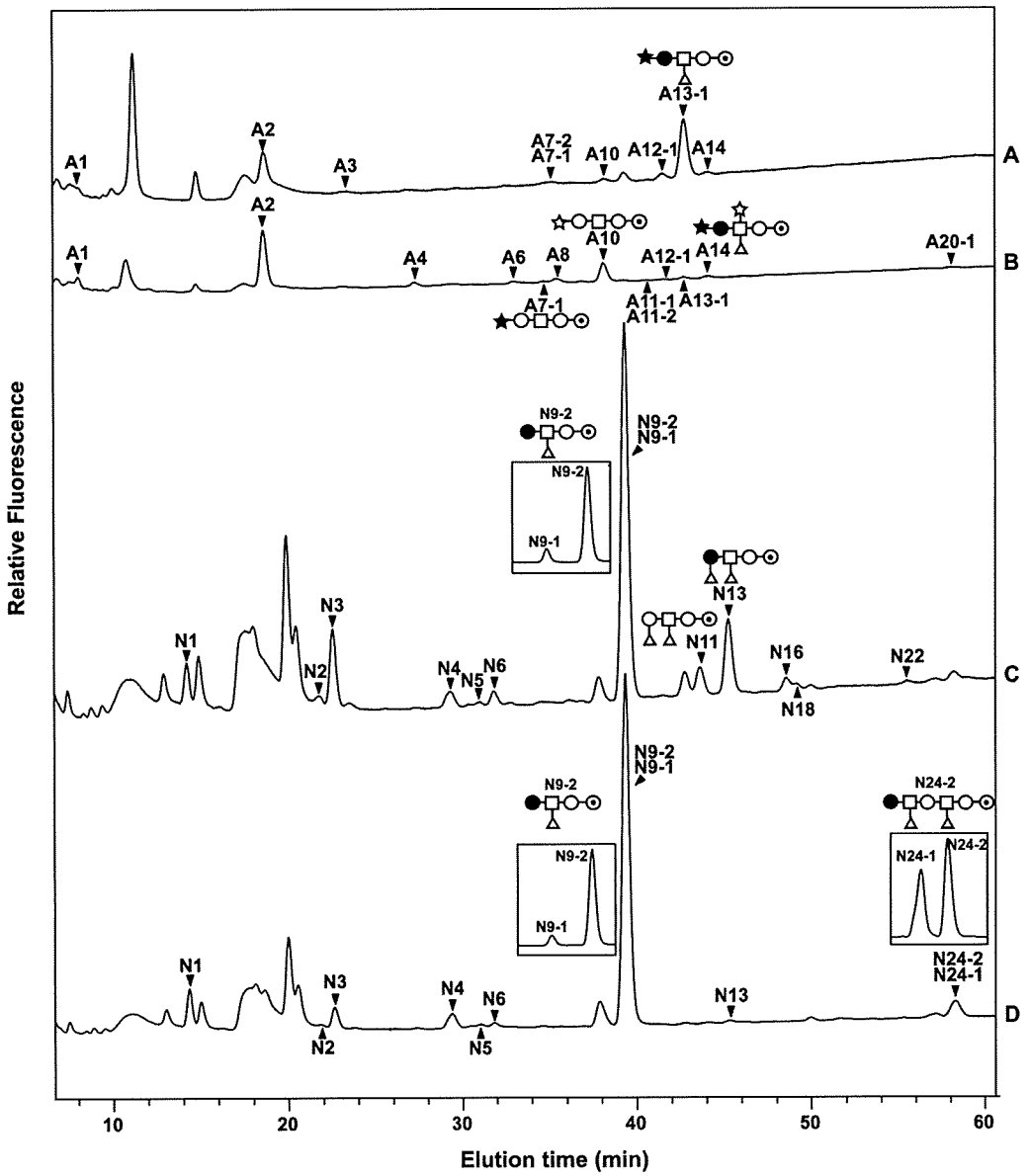

Figure 5. Comparison of amide column HPLC profiles of acidic and neutral PA-oligosaccharide obtained from CCs and NCs from case 15. Specific elevation of α2−3 sialylation of Lc_4 in carcinogenesis. (A) Acidic fraction of CCs, (B) acidic fraction of NCs, (C) neutral fraction of CCs, (D) neutral fraction of NCs. Identified PA-oligosaccharides of each peak, the ratio of mixtures of peak N9 and N24, and schemes are shown as in Figure 2.

addition, significant amounts of Lc_6 (N12), $VI^2Fuc\alpha$-Lc_6 (N17−1) and $V^3Fuc\alpha$-$_{1,2}Lc_6$ (N19) are found (Figure 6D). The structures of CCs can be basically explained by the three different elevated levels previously described in conjunction with the structures of NCs as described above. In the neutral fraction from CCs, peaks N9 and N24 are composed of only Le^x (N9−1) and $V^3Fuc\alpha III^3Fuc\alpha$-$nLc_6$ (N24−1), respectively (Figure 6C). Le^y (N11) and nLc_4 (N6) are increased, and Lc_4 (N5) is decreased due to the elevation of type-2 chain oligosaccharides. In addition, agalacto-$III^3Fuc\alpha$-nLc_6 (N14), nLc_6 (N15), $III^3Fuc\alpha$-nLc_6 (N21) and $VI^2Fuc\alpha V^3Fuc\alpha$-$nLc_6$ (N22) appeared as minor peaks in CCs. In the acidic fraction from CCs (Figure 6A), LST-c (A10) is increased and $IV^2Fuc\alpha IV^6NeuAc\alpha$-$nLc_4$(A12−2), $VI^6NeuAc\alpha III^3Fuc\alpha$-$nLc_6$ (A18) and $VI^6NeuAc\alpha$-$_{1,2}Lc_6$ (A16−2) appeared as the results of elevated α2−6 sialylation. Significant amounts of SLe^x (A12−1) were observed, as the result of elevated α2−3 sialylation. SLe^a (A13−1), which was found in almost all the CCs, was not found due to the lack of α1−4 fucosyltransfarase activity, but instead, LST-a (A7−2), nonfucosylated SLe^a, was observed as one of the major peaks.

Comparison of Activities of Sialyltransferase, Fucosyltransferase, and β-Galactosyltransferase from CCs and NCs. To investigate whether the three types of alteration in glycosylation observed in oncogenesis depend on changes in the activities of related glycosyltransferases, β-galactosyltransferase, sialyltransferase and fucosyltransefarse from CCs and NCs were examined (Figure 7). Measurement of the activities from CCs and NCs from 2 of the 16 cases (case5 and 10) were not performed since the amount of sample was lacking. In terms of increase of type-2 chain oligosaccharides, β1−3 and β1−4 galactosyltranferase activities using Lc_3-PA as an acceptor were examined (Figure 7A and B). As shown in Figure 7A, high levels of β1−3 galactosyltransferase activity (more than 10 000 pmol/mg protein/hour) were found in NCs from 10 cases, and low levels of β1−3 galactosyltranferase activities (less than 10 000 pmol/mg protein/hour) were found in NCs from 4 cases (case

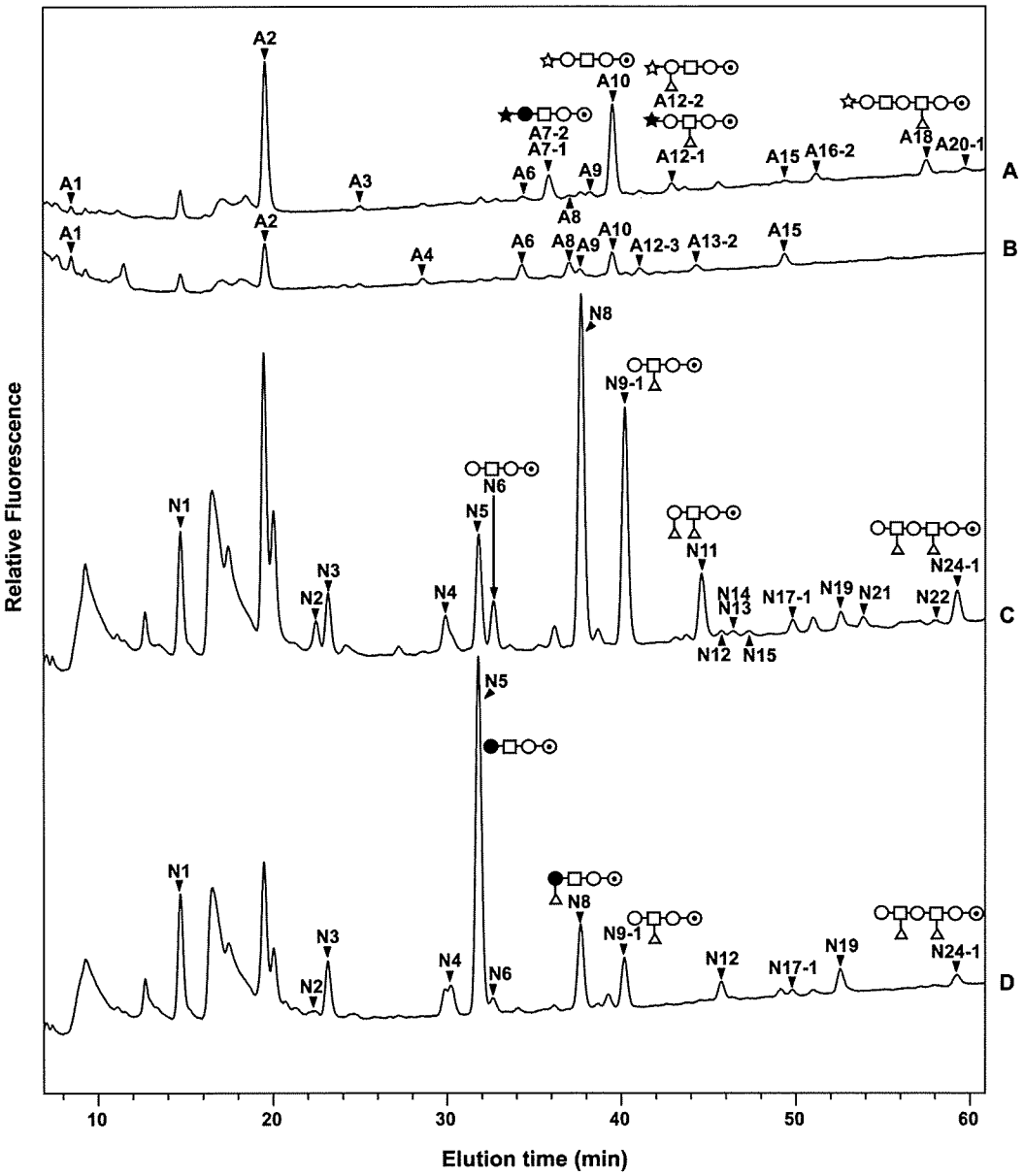

Figure 6. Comparison of amide column HPLC profiles of acidic and neutral PA-oligosaccharide obtained from CCs and NCs from case 16, who lacks Lewis enzyme activity. (A) Acidic fraction of CCs, (B) acidic fraction of NCs, (C) neutral fraction of CCs, (D) neutral fraction of NCs. Identified PA-oligosaccharides of each peak and schemes are shown as in Figure 2.

7, 8, 9, and 12). The β1−3 galactosyltransferase activities were markedly decreased or at sustained low levels in malignant transformation (Figure 7A). In contrast to the β1−3 galactosyltranferase activities, β1−4 galactosyltransferase activities were found in all NCs with little variation. Significant differences between the activity in CCs and NCs were not observed in any of the cases (Figure 7B).

Four kinds of sialyltransferase activity were examined. The sialyltransferase activities of α2−6 to terminal galactose of type-2 chains to generate LST-c from nLc_4 was observed in NCs from all cases and increased in carcinogenesis in all cases with 2 exceptions (cases 14 and 15) (Figure 7C). The two exceptional cases, case 14 and 15, represent specific elevation of α2−3 sialylation together with down regulation of α2−6 sialylation in malignant transformation, as shown in Figure 5. The amounts of sialic acid linked α2−6 to subterminal GlcNAc of type-1 chains are decreased or at sustained low levels in malignant transformation, which is also similar to the alteration found in the β1−3 galactosyltranferase activities (Compare Figure 7A and E). The changes in the activities giving sialic acid transfer α2−3 to terminal galactose of type 2 and type 1 lactosamine chains in malignant transformation vary case to case, with some increasing, some decreasing and others showing no change (Figure 7D and F). Relatively high α1−2 fucosyltranferase activity toward terminal galactose of nLc_4 and Lc_4 were observed in NCs from 3 cases (cases 8, 9, 12, Figure 7G, H). Low to negligible levels of α1−2 fucosyltranferase activities were found in NCs from other 11 cases (Figure 7G, H). Profound change, either increased or decreased, in α1−2 fucosyltranferase activity occurred in carcinogenesis. When α1−2 fucosyltranferase activities are very low or at negligible levels in NCs, marked increases in the activities were found in CCs. In contrast, when α1−2 fucosyltranferase activities are high in NCs, marked decreases in the activities were found in CCs (case 9). Very high levels of α1−3 and α1−4

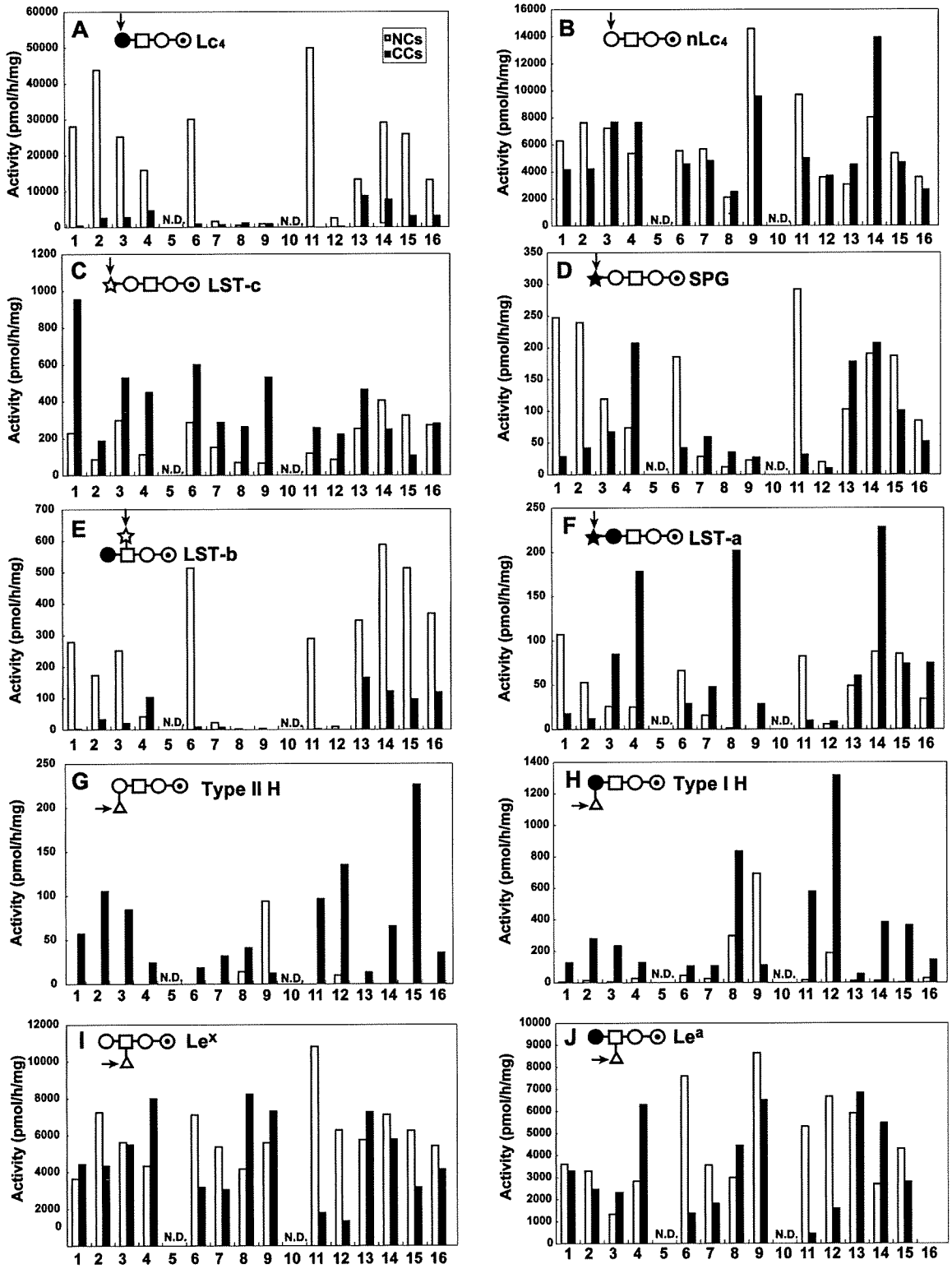

Figure 7. Comparison of activities of β-galactosyltransferases, sialyltransferases, and fucosyltransferases in CCs and NCs. Open and filled bars indicate NCs and CCs, respectively. (A and B) Activities of β1−3 and β1−4 galactosyltransferases using Lc_3-PA as acceptor. (C and D) Activities of sialic acid transfer α2−6 and α2−3 to the nonreducing terminal galactose of type-2 lactosamine, using nLc_4-PA as acceptor. (E and F) Activities of sialic acid transfer α2−6 to subterminal GlcNAc and α2−3 to nonreducing terminal galactose of type-1 lactosamine, using Lc_4-PA as acceptor. (G and H) Activities of α1−2 fusosyltransferases, using nLc_4-PA and Lc_4-PA as acceptors. (I and J) Activities of α1−3 and α1−4 fusosyltransferases, using nLc_4-PA and Lc_4-PA as acceptors, respectively. Assays conditions were as described in Experimental Procedures. N.D., Not determined due to the lack of remaining samples. Schemes and abbreviations of reaction products are shown in each part. Oligosaccharides linked to the reaction are highlighted with arrows.

fucosyltransferase activities toward subterminal GlcNAc of nLc_4 and Lc_4, respectively, were found in all NCs, and alteration of the activities varies from decrease to increase in malignant transformation (Figure 7I, J). As mentioned above, α1−4 fucosyltransferase activities were not found at any extent in NCs and CCs from case 16 (Figure 7J).

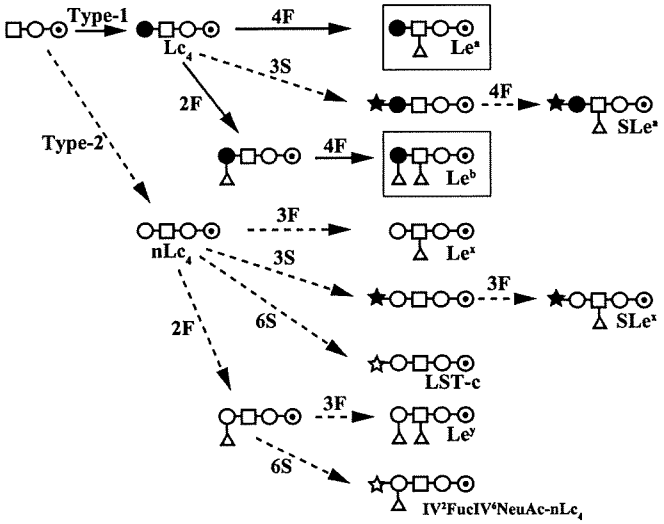

Figure 8. Synthetic pathways for major groups of GSLs in CCs and NCs. Arrows indicate the pathways predominating in NCs. Broken arrows indicate the pathways increased in carcinogenesis of CCs. Abbreviations: 4F, α1−4 fucosylation of GlcNAc, 3F, α1−3 fucosylation of GlcNAc, 2F, α1−2 fucosylation of galactose, 3S, α2−3 sialylation of galactose, 6S, α2−6 sialylation of galactose. Schemes are shown as in Figure 2. The structures of GSLs in NCs are composed of mainly Le^a and Le^b (highlighted by square), formed by α1−4 fucosylation of Lc_4, and α1−2 fucosylation of Lc_4 followed by α1−4 fucosylation, respectively. In malignant transformation the type-2 ratio, α2−3 and/or α2−6 sialylation, and α1−2 fucosylation are increased. These alterations result in increases in the amounts of or the appearance of oligosaccharides such as Le^x, Le^y, LST-c, SLe^x, SLe^a and IV^2FucIV^6NeuAc-nLc_4.

Discussion

The accumulation of GSLs having type-2 chain derivatives, i.e. those with Le^x, Le^y, dimeric Le^x, and their sialosyl derivatives in colon cancers as analyzed by conventional methods has been reported.[7−9] Furthermore, the accumulation of GSLs having α2−6 sialylated lactosaminyl structures in human colon cancers has been detected using specific monoclonal antibodies and α2−6 sialyl specific lectins.[21−25] Although most of the previous findings are essentially in agreement with our results, we provide much more detailed, precise structural information in terms of quality and quantity of the GSLs from both NCs and CCs from many patients.

We found three kinds of changes in oligosaccharide structures in malignant transformation of CCs. Synthetic pathways for the major groups of GSLs in CCs and NCs are outlined in Figure 8. However, these apparent alterations are more difficult to identify when whole cancerous tissues and normal tissues are used as analyzing sources (data not shown). This result indicates the importance of the isolation of cells with high purity for cancer glycomic analyses.

Metastasis to the liver is the most important factor of poor prognosis of colorectal cancers. Even though extensive studies have been performed in attempts to elucidate the molecular mechanism of this event, a different approach may help to better understand the mechanism or help discover promising predictive factors clinically undetected during initial surgery.[26] In this study, we found a common feature of the structures of the GSLs from CCs of the 5 patients: namely a marked elevation of type-2 oligosaccharides. One hypotheses as to how the alteration is associated with hepatic metastasis can be considered; the presence of an oligosaccharide determinant involved in hepatic metastasis. SLe^x and SLe^a are thought to function as E-selectin ligands and to be involved in hematogenous metastasis of cancers.[4] However, when considering the structures of GSLs, SLe^x and SLe^a determinants would seem unlikely to be critically involved in hepatic metastasis of CCs because expression of SLe^x is lacking in one of the five patients and SLe^a is absent in another. However, in contrast to our results, increased expression of SLe^x and SLe^a in colon cancers was found to be correlated with poor prognosis in patients with colorectal cancers by clinico-immunohistochemical analyses using paraffin embedded colon cancer tissues.[27,28] However, the results from the previous immunohistochemical analyses did not reflect the quantities of GSLs, because we found that GSLs are removed by alcohol dehydration and xylene treatment in the process of embedding in paraffin. The most probable candidate oligosaccharide determinant of GSLs involved in liver metastasis is thought to be α2−6 sialylated $IV^2Fucα$-nLc_4 (A12−2, Figures 4, 6). This GSL was isolated and the structures identified in our previous study which analyzed the structures of GSLs from colon cancer tissue at hepatic metastasis.[17] Marked elevation of nLc_4, followed by α1−2 fucosylation and α2−6 sialylation of terminal galactose, results in the generation of this structure (Figure 8). The enzymes responsible and the reaction mechanism that generates this unique structure have been already investigated and submitted elsewhere. This GSL was found in the CCs of all 5 patients having hepatic metastasis and 4 other patients in whom liver metastasis had not been shown, out of a total of 16 subjects.

Cancer malignancy is defined by several key phenotypes, including apoptosis, motility, angiogenesis, self-adhesion, adhesion to extracellular matrix and to endothelial cells and aberrant glycosylation is thought to be involved in these steps[3]. It is unclear at present as to which of the above processes the altered GSL structures on the surface of CCs found in this study are involved. However, when focusing on hepatic metastasis, functional analyses of α2−6 sialylated $IV^2Fucα$-nLc_4 may help to solve the problem. To this end, a study looking into this line of investigation is under way.

Precise analyses of the activities of glycosyltransferases responsible for the aberrant glycosylation in malignant transformation presented us with valuable information to help us understand the mechanisms involved. Three types of increases in levels, the ratio of type-2 oligosaccharides, α2−6 sialylation and α1−2 fucosylation can be approximately accounted for by changes in the activities of related glycosyltransferases. Thus, in malignant transformation, activities of β1−3 galactosyltransferase were markedly decreased, the activities of α2−6 sialyltransferase toward terminal galactose of nLc4 are increased, and the activities of α1−2 fucosyltransferase toward nLc_4 and Lc_4 are markedly increased with no or a few exceptions. It is possible that greatly reduced activity of β1−3 galactosyltransferase and a virtually invariant alteration in the activities of β1−4 galactosyltransferase in carcinogenesis result in the increase of type-2 chain oligosaccharides. Similarly, increased type-2 chain oligosaccharides, followed by the increase in the activity of α2−6 sialyltransferase toward type-2 lactosamine chains results in the elevation of α2−6 sialylated type-2 oligosaccahrides, such as LST-c and $VI^6NeuAcαIII^3Fucα$-nLc_6. Furthermore, greatly increased activity of α1−2 fucosyltransferase toward both nLc_4 and Lc_4 leads to the elevation of α1−2 fucosylated products, such as Le^y and Le^b. In contrast, elevation of α2−3 sialylation in carcinogenesis does not depend on

changes in the related activities. For example, levels of Lc_4 in NCs were higher than in CCs in almost all the cases. In addition, in NCs the levels of the two glycosyltransferase activities that generate SLe^a from Lc_4 ($\alpha2-3$ sialyltransferase and $\alpha1-4$ fucosyltransferase), are similar to or higher than those in CCs in several cases, (e.g., cases 1, 2, 6, 11, 12, 13, 15). However, expression of SLe^a is observed in CCs, but absent or present in only very small quantities in NCs. This maybe a result of the concerted actions of glycosyltransferases. To generate the SLe^a structure, it is essential that $\alpha2-3$ sialyltransferase acts before $\alpha1-4$ fucosyltransferase, because $\alpha2-3$ sialyltransferase does not act on terminal galactose when the adjacent GlcNAc is fucosylated. Hence, a well arranged mechanism of preferential $\alpha1-4$ fucosylation on type-1 oligosaccharides, such as Lc_4, exists in NCs but breaks down in CCs.

The structures of oligosaccharides analyzed in this study were limited to those of GSLs. However, other glycans, such as N- and O-linked glycoproteins and proteoglycans, are also thought to be intimately involved in cancer malignancy.[3,29] It is therefore also important to pursue the structures and functional roles of these glycans on the surface of cancer cells.

This clinico-glycomic study revealed three kinds of unidirectional changes in glycosylation in carcinogenesis of CCs and examined the activities of related glycosyltransferases. Because the number of cases analyzed is small, in order to be able to generalize about the observations in this study, a larger number of samples is required. However, the findings from this study will give important clues toward the elucidation of the detailed mechanism of alteration of glycosylation and its involvement in cancer malignancy.

Acknowledgment. This work was supported in part by research grants from the Ministry of Health, Labour and Welfare of Japan. We thank Drs. Koichi Honke and Shunji Natsuka for useful discussions and critical comments on the manuscript.

Supporting Information Available: Supplemental Table 1. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Hakomori, S. Aberrant glycosylation in tumors and tumor-associated carbohydrate antigens. *Adv. Cancer Res.* **1989**, *52*, 257–331.
(2) Hakomori, S. Tumor malignancy defined by aberrant glycosylation and sphingo(glyco)lipid metabolism. *Cancer Res.* **1996**, *56* (23), 5309–18.
(3) Hakomori, S. Glycosylation defining cancer malignancy: new wine in an old bottle. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (16), 10231–3.
(4) Kannagi, R.; Izawa, M.; Koike, T.; Miyazaki, K.; Kimura, N. Carbohydrate-mediated cell adhesion in cancer metastasis and angiogenesis. *Cancer Sci.* **2004**, *95* (5), 377–84.
(5) Siddiqui, B.; Whitehead, J. S.; Kim, Y. S. Glycosphingolipids in human colonic adenocarcinoma. *J. Biol. Chem.* **1978**, *253* (7), 2168–75.
(6) Hakomori, S.; Nudelman, E.; Levery, S. B.; Patterson, C. M. Human cancer-associated gangliosides defined by a monoclonal antibody (IB9) directed to sialosyl alpha 2 leads to 6 galactosyl residue: a preliminary note. *Biochem. Biophys. Res. Commun.* **1983**, *113* (3), 791–8.
(7) Hakomori, S.; Nudelman, E.; Levery, S. B.; Kannagi, R. Novel fucolipids accumulating in human adenocarcinoma. I. Glycolipids with di- or trifucosylated type 2 chain. *J. Biol. Chem.* **1984**, *259* (7), 4672–80.
(8) Fukushi, Y.; Nudelman, E.; Levery, S. B.; Hakomori, S.; Rauvala, H., III. A hybridoma antibody (FH6) defining a human cancer-associated difucoganglioside (VI3NeuAcV3III3Fuc2nLc6). *J. Biol. Chem.* **1984**, *259* (16), 10511–7.
(9) Nudelman, E.; Levery, S. B.; Kaizu, T.; Hakomori, S. Novel fucolipids of human adenocarcinoma: characterization of the major Ley antigen of human adenocarcinoma as trifucosylnonaosyl Ley glycolipid (III3FucV3FucVI2FucnLc6). *J. Biol. Chem.* **1986**, *261* (24), 11247–53.
(10) Nudelman, E.; Fukushi, Y.; Levery, S. B.; Higuchi, T.; Hakomori, S. Novel fucolipids of human adenocarcinoma: disialosyl Lea antigen (III4FucIII6NeuAcIV3NeuAcLc4) of human colonic adenocarcinoma and the monoclonal antibody (FH7) defining this structure. *J. Biol. Chem.* **1986**, *261* (12), 5487–95.
(11) Saito, S.; Orikasa, S.; Ohyama, C.; Satoh, M.; Fukushi, Y. Changes in glycolipids in human renal-cell carcinoma and their clinical significance. *Int. J. Cancer* **1991**, *49* (3), 329–34.
(12) Korekane, H.; Shida, K.; Murata, K.; Ohue, M.; Sasaki, Y.; Imaoka, S.; Miyamoto, Y. Evaluation of laser microdissection as a tool in cancer glycomic studies. *Biochem. Biophys. Res. Commun.* **2007**, *352* (3), 579–86.
(13) Momburg, F.; Moldenhauer, G.; Hammerling, G. J.; Moller, P. Immunohistochemical study of the expression of a Mr 34,000 human epithelium-specific surface glycoprotein in normal and malignant tissues. *Cancer Res.* **1987**, *47* (11), 2883–91.
(14) Ellis, W. J.; Pfitzenmaier, J.; Colli, J.; Arfman, E.; Lange, P. H.; Vessella, R. L. Detection and isolation of prostate cancer cells from peripheral blood and bone marrow. *Urology* **2003**, *61* (2), 277–81.
(15) Kemmner, W.; Moldenhauer, G.; Schlag, P.; Brossmer, R. Separation of tumor cells from a suspension of dissociated human colorectal carcinoma tissue by means of monoclonal antibody-coated magnetic beads. *J. Immunol. Methods* **1992**, *147* (2), 197–200.
(16) Kruger, W.; Datta, C.; Badbaran, A.; Togel, F.; Gutensohn, K.; Carrero, I.; Kroger, N.; Janicke, F.; Zander, A. R. Immunomagnetic tumor cell selection--implications for the detection of disseminated cancer cells. *Transfusion* **2000**, *40* (12), 1489–93.
(17) Korekane, H.; Tsuji, S.; Noura, S.; Ohue, M.; Sasaki, Y.; Imaoka, S.; Miyamoto, Y. Novel fucogangliosides found in human colon adenocarcinoma tissues by means of glycomic analysis. *Anal. Biochem.* **2007**, *364* (1), 37–50.
(18) Ito, M.; Yamagata, T. Purification and characterization of glycosphingolipid-specific endoglycosidases (endoglycoceramidases) from a mutant strain of Rhodococcus sp. Evidence for three molecular species of endoglycoceramidase with different specificities. *J. Biol. Chem.* **1989**, *264* (16), 9510–9.
(19) Natsuka, S.; Hase, S. Analysis of N- and O-glycans by pyridylamination. *Methods Mol. Biol.* **1998**, *76*, 101–13.
(20) Tokugawa, K.; Oguri, S.; Takeuchi, M. Large scale preparation of PA-oligosaccharides from glycoproteins using an improved extraction method. *Glycoconj. J.* **1996**, *13* (1), 53–6.
(21) Hakomori, S.; Patterson, C. M.; Nudelman, E.; Sekiguchi, K. A monoclonal antibody directed to N-acetylneuraminosyl-alpha 2 leads to 6-galactosyl residue in gangliosides and glycoproteins. *J. Biol. Chem.* **1983**, *258* (19), 11819–22.
(22) Nilsson, O.; Lindholm, L.; Holmgren, J.; Svennerholm, L. Monoclonal antibodies raised against NeuAc alpha 2−6neolactotetraosylceramide detect carcinoma-associated gangliosides. *Biochim. Biophys. Acta* **1985**, *835* (3), 577–83.
(23) Sata, T.; Roth, J.; Zuber, C.; Stamm, B.; Heitz, P. U. Expression of alpha 2,6-linked sialic acid residues in neoplastic but not in normal human colonic mucosa. A lectin-gold cytochemical study with Sambucus nigra and Maackia amurensis lectins. *Am. J. Pathol.* **1991**, *139* (6), 1435–48.
(24) Dall'Olio, F.; Trere, D. Expression of alpha 2,6-sialylated sugar chains in normal and neoplastic colon tissues. Detection by digoxigenin-conjugated Sambucus nigra agglutinin. *Eur. J. Histochem.* **1993**, *37* (3), 257–65.
(25) Yamashita, K.; Fukushima, K.; Sakiyama, T.; Murata, F.; Kuroki, M.; Matsuoka, Y. Expression of Sia alpha 2→6Gal beta 1→4GlcNAc residues on sugar chains of glycoproteins including carcinoembryonic antigens in human colon adenocarcinoma: applications of Trichosanthes japonica agglutinin I for early diagnosis. *Cancer Res.* **1995**, *55* (8), 1675–9.
(26) Rudmik, L. R.; Magliocco, A. M. Molecular mechanisms of hepatic metastasis in colorectal cancer. *J. Surg. Oncol.* **2005**, *92* (4), 347–59.
(27) Nakamori, S.; Kameyama, M.; Imaoka, S.; Furukawa, H.; Ishikawa, O.; Sasaki, Y.; Kabuto, T.; Iwanaga, T.; Matsushita, Y.; Irimura, T. Increased expression of sialyl Lewisx antigen correlates with poor

survival in patients with colorectal carcinoma: clinicopathological and immunohistochemical study. *Cancer Res.* **1993**, *53* (15), 3632–7.

(28) Nakayama, T.; Watanabe, M.; Katsumata, T.; Teramoto, T.; Kitajima, M. Expression of sialyl Lewis(a) as a new prognostic factor for patients with advanced colorectal carcinoma. *Cancer* **1995**, *75* (8), 2051–6.

(29) Zhao, Y. Y.; Takahashi, M.; Gu, J. G.; Miyoshi, E.; Matsumoto, A.; Kitazume, S.; Taniguchi, N. Functional roles of N-glycans in cell signaling and cell adhesion in cancer. *Cancer Sci.* **2008**, *99* (7), 1304–10.

PR900092R

Using gene expression profiling to identify a prognostic molecular spectrum in gliomas

Mitsuaki Shirahata,[1,2] Shigeyuki Oba,[3] Kyoko Iwao-Koizumi,[2] Sakae Saito,[2] Noriko Ueno,[2] Masashi Oda,[1] Nobuo Hashimoto,[1] Shin Ishii,[3] June A. Takahashi[1] and Kikuya Kato[2,4]

[1]Department of Neurosurgery, Kyoto University Graduate School of Medicine, 54 Kawaharacho, Shogoin Sakyoku, Kyoto, 606-8507, Japan; [2]Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases, 1-3-3 Nakamichi, Higashinari-ku, Osaka, 537-8511, Japan; [3]Laboratory for Theoretical Life Science, Nara Institute of Science and Technology 816-5 Takayama, Ikoma, Nara, 630-0192, Japan

(Received April 29, 2008/Revised August 7, 2008; September 1, 2008/Accepted September 12, 2008/Online publication November 25, 2008)

Histopathological classification of gliomas is often clinically inadequate due to the diversity of tumors that fall within the same class. The goal of the present study was to identify prognostic molecular features in diffusely infiltrating gliomas using gene expression profiling. We selected 3456 genes expressed in gliomas, including 3012 genes found in a glioma expressed sequence tag collection. The expression levels of these genes in 152 gliomas (100 glioblastomas, 21 anaplastic astrocytomas, 19 diffuse astrocytomas, and 12 anaplastic oligodendrogliomas) were measured using adapter-tagged competitive polymerase chain reaction, a high-throughput reverse transcription–polymerase chain reaction technique. We applied unsupervised and supervised principal component analyses to elucidate the prognostic molecular features of the gliomas. The gene expression data matrix was significantly correlated with the histological grades, oligo-astro histology, and prognosis. Using 110 gliomas, we constructed a prediction model based on the expression profile of 58 genes, resulting in a scheme that reliably classified the glioblastomas into two distinct prognostic subgroups. The model was then tested with another 42 tissues. Multivariate Cox analysis of the glioblastoma patients using other clinical prognostic factors, including age and the extent of surgical resection, indicated that the gene expression profile was a strong and independent prognostic parameter. The gene expression profiling identified clinically informative prognostic molecular features in astrocytic and oligodendroglial tumors that were more reliable than the traditional histological classification scheme. (*Cancer Sci* 2009; 100: 165–172)

Despite being critical for treatment outcomes, precisely assessing the risk of a glioma using histological classification fails to address the heterogeneity of responses to therapy among patients within the same histological class, indicating that the classification system is not an adequate predictor of the clinical behavior of the tumor.[1] However, recent studies suggest that molecular approaches are useful for identifying prognostic markers. Genetic analyses have shown that allelic loss of chromosomes 1p and 19q is a strong predictor of longer survival in patients with oligodendroglial tumors.[2] Furthermore, MGMT promoter methylation was found to be an independently favorable prognostic factor in GB patients.[3]

In addition to genomic changes in glioma cells, gene expression profiling is expected to elucidate molecular features related to clinical parameters. Application of this method to glioma patients will not only help clinicians make an optimal clinical decision, but also lead to possibilities for personalized, pathway-targeted therapies in the future.

In the present report, we describe high-throughput RT-PCR-based gene expression profiling of more than 150 gliomas. Previously we established a molecular diagnostic system for AO and GB, using part of the data matrix.[4] We extended the study, and identified specific gene expression patterns that can be used to classify GB into two distinct subgroups, which are strongly predictive of prognosis.

Materials and Methods

Patient characteristics. We obtained 169 glioma specimens from patients who underwent surgical resection at Kyoto University Hospital or nearby regional hospitals between 1998 and 2005. The majority of the patients were recruited from a phase II clinical trial (the KNOG study).[5] The protocol for the present study was approved by the Institutional Review Board of Kyoto University. Written informed consent was obtained from each of the patients.

The following samples were excluded from the study: samples with evidence of previous chemoradiation therapy within the last 10 years (five samples), insufficient tumor content (four samples), poor expression data quality (two samples), or the presence of pilocytic astrocytoma (six samples). In total, 152 samples, including 100 GB, 21 AA, 19 DA, and 12 AO, were eligible for further analysis. The specimens were examined histologically at the primary hospitals according to the World Health Organization's 2000 criteria.[6] The original slides were reviewed by the Kyoto University Pathology Unit for the final diagnosis. The samples were collected at the time of the initial surgery without any prior treatment, except for one sample of recurrent AO collected 14 years after the initial treatment, which included radiotherapy. The preoperative Karnofsky performance status score was at least 50 for each of the cases. The extent of surgical resection was classified into one of three categories according to postoperative MRI carried out shortly after surgery: complete resection, incomplete resection, or biopsy. Complete resection was defined as no evidence of enhanced lesion or T1 abnormality in non-enhanced tumors using postoperative MRI. All patients received fractionated local radiotherapy with or without ACNU-based chemotherapy, except for one case of diffuse astrocytoma that was not treated with any adjuvant therapy, and one case of recurrent AO that was treated with adjuvant chemotherapy alone. Seventy-three of the 100 GB patients were treated according to the regimen from the KNOG study. For tumor progression, the patients underwent a second operation if possible, and received further chemotherapy for most of the cases.

RNA and DNA isolation. In all cases, tumor specimens were dissected into two portions at surgery, one for histological

[4]To whom correspondence should be addressed. E-mail: katou-ki@mc.pref.osaka.jp

Abbreviations: AA, anaplastic astrocytoma; ACNU, numustine; AO, anaplastic oligodendroglioma; ATAC-PCR, adaptor-tagged competitive polymerase chain reaction; DA, diffuse astrocytoma; EST, expressed sequence tag; GB, glioblastoma; KNOG, Kyoto Neuro-Oncology Group; LDA, linear discriminant analysis; MDA, M.D. Anderson Cancer Center; MGH, Massachusetts General Hospital; MGMT, O^6-methyl guanine methyltransferase; MRI, magnetic resonance imaging; OS, overall survival; PC1, first principal component; PCA, principal component analysis; PFS, progression-free survival; RT-PCR, reverse transcription–polymerase chain reaction.

doi: 10.1111/j.1349-7006.2008.01002.x

diagnosis and the other for molecular experiments. The tumor specimens for molecular research were immediately snap frozen at surgical resection, and kept at –80°C until use. Total RNA was extracted from 100 mg of the tumor specimen with Trizol reagent (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. Tumor specimens containing 20% or more of non-tumor or necrotic areas were excluded from further analysis. Genomic DNA was isolated utilizing a QIAamp DNA Mini Kit (Qiagen, Valencia, CA, Germany) according to the manufacturer's instructions.

Gene expression analysis. The expression of genes was measured by ATAC-PCR. Selection of the genes examined was based on an EST sequencing survey of the genes expressed in 12 glioma tissues as described previously.[7] We identified 3012 unique sequences from the EST collection, and prepared 3456 primers for ATAC-PCR, including primers to survey an additional 444 genes selected from a literature survey. The ATAC-PCR procedure has been described previously.[8] The complete list of genes and expression data from the present study are shown in a supporting file. The expression data will be deposited into Center for Information Biology gene EXpression database (CIBEX) at the DNA Data Bank of Japan.

Methylation-specific PCR. DNA methylation patterns in the CpG islands of the MGMT gene were determined as described previously.[3,9]

Statistical methods.

Survival data. OS and PFS were defined as the period from surgery to death and that from surgery to radiological detection of tumor progression, respectively. Tumor progression was evaluated according to the criteria of the committee of the Brain Tumor Registry (Japan): a 25% increase in tumor size, appearance of new lesions, or obvious deterioration due to mass effects or perifocal edema.[10] A radiological examination was carried out every 2 or 3 months postoperatively in high-grade glioma patients, and immediately after neurological deterioration was detected.

Data preparation and preprocessing. The raw expression data were first adjusted to correct for variations due to different sample mRNA concentrations by dividing each value by the corresponding median value. Values less than 0.05 and more than 20 were converted to 0.05 and 20, respectively. The entire data matrix was then converted to a logarithmic scale. Genes for which 20% of the data were missing were excluded from statistical analysis after missing value imputation using BPCAfill.[11] The gene expression levels were then normalized so that the genewise mean for each sample became zero.

For the analysis with external data sets, we obtained data for 24 genes from the MGH data set[12] and 55 genes from the MDA data set[13] for the profile of 58 genes. After a logarithmic conversion, normalization was carried out so that the average gene expression level for each gene was zero. Zero was used as the value when data were not available.

Feature extraction. We used two feature-extraction methods to obtain effective coordinate axes onto which each data vector could be projected appropriately. Unsupervised principal component analysis (PCA) was used to extract axes (principal axes) representing variations in sample expression vectors. Because the vector dimensionality (i.e. the number of genes) was larger than the number of vectors (i.e. the number of patients), we carried out singular value decomposition to obtain the principal axes.[14]

The second method was LDA, which searches for an axis on which the signal-to-noise ratio between the projected data and the biological and clinical labels of interest is maximized. In the present study, LDA was carried out after the gene expression matrix had been projected onto a two-dimensional principal component space generated by unsupervised PCA. The combination of LDA and PCA is also known as a principal component regression, which rarely over-fits when dimensionality is reduced enough by the PCA process. A low-dimensional relationship between the samples and the biological and clinical labels was obtained.

Cox proportional hazards regression. To select genes for the prognosis prediction model, for which the target was PFS (see supervised PCA below), we evaluated the significance of the correlation with PFS using univariate Cox tests (the log-rank test).

In addition, multivariate Cox proportional hazard regression was used to calculate regression coefficients between possible prognostic genes and PFS of the corresponding patients; the obtained coefficient vector represented an axis with coordinates that showed the strongest correlation with PFS. As with the LDA, calculations were carried out after the application of PCA, which extracted two-dimensional representations. Cox's analyses were done by our original Matlab 6.5 implementation that mimicked Cox's analysis modules in XploRe (http://www.xplore-stat.de/), a web-based statistics software.

Among the GB patients, the significance of the final molecular classification compared to the other prognostic factors was evaluated using multivariate Cox analysis. We included both the training and test sets for the analyses.

Supervised PCA. When constructing the prognosis prediction model whose target was PFS, we used supervised PCA, in which genes correlated with a variable of interest (here, PFS) were selected first. Supervised PCA was then carried out in the subspace represented by the selected genes. The resulting PC1 score exhibited a strong correlation with the variable.[15] In the present study, the genes correlated with PFS were selected using Cox proportional hazards regression.

Results

We carried out survival analysis on 152 tissue samples (100 GB, 21 AA, 19 DA, and 12 AO). Because the survival benefits of various chemotherapeutic regimens for glioma, especially GB, are not distict,[16–19] with the exception of that associated with temozolomide,[20,21] we included all of the cases in the survival analysis irrespective of the chemotherapy treatment. After data processing, we obtained a data matrix consisting of 3225 genes from the 152 samples. We divided the data matrix into two – one set consisted of 110 patients (the training set) and the other contained 42 patients (the test set) – by selecting samples that arrived at the laboratory at earlier dates for the training set.

The samples that were used as the training set consisted of 77 GB, 11 AA, 11 DA, and 11 AO. The median age at surgery of the associated patients was 54 years (range 21–82 years). The median follow-up period for the survivors with GB was 19.5 months (range 3–62 months). Among the 77 GB patients, 61 patients showed tumor progression and 48 patients died. The median PFS and OS periods in the GB patients were 7 and 14 months, respectively. Data about OS is generally more accurate than that for PFS. However, because the PFS data were carefully obtained following the strict guidelines of the KNOG study,[5] the quality of the PFS data was comparable to that of the OS data. Because OS may be affected by treatment bias at the time of tumor progression, such as a second operation, and there was a good correlation between OS and PFS (correlation coefficient 0.96), we adopted PFS as the clinical parameter that most accurately represented the aggressiveness of the gliomas in each patient.

Unsupervised PCA of the training data set revealed that the cumulative contribution ratios of the top six principal components were 0.9076, 0.9500, 0.9595, 0.9673, 0.9719, and 0.9773. We plotted the samples on a plane constructed using the first two components. The resulting scatter diagram demonstrated that the sample distributions were related to the histological classes, indicating a close correlation between the global gene expression patterns and the histology of the samples, such as malignancy grades and oligo-astrocytic characteristics. Interestingly,

doi: 10.1111/j.1349-7006.2008.01002.x

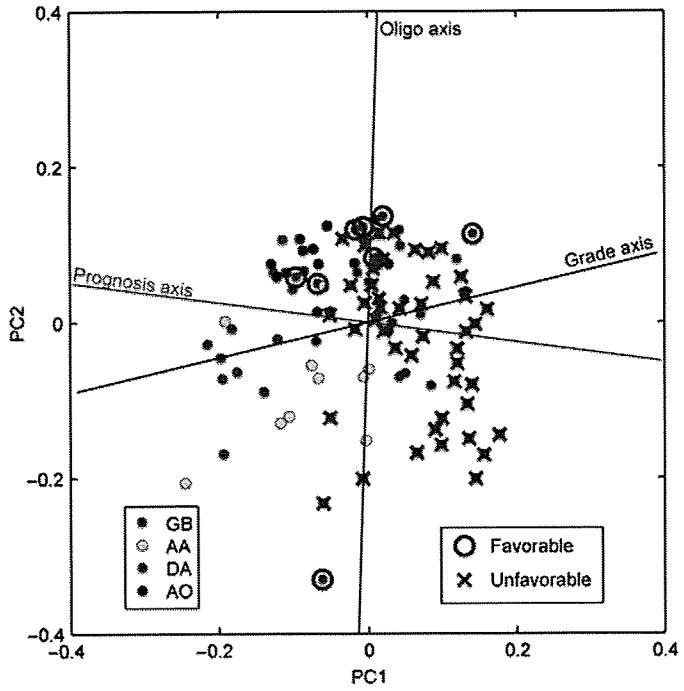

Fig. 1. Principal component analysis based on the expression of 3225 genes in 110 gliomas in the training set. Circles indicate tumor samples with color representing histological classes. Glioblastoma (GB) samples with progression-free survival of 2 years or longer (favorable) are circled in blue, and those with overall survival shorter than 2 years (unfavorable) are marked with a red 'X'. AA, anaplastic astrocytoma; AO, anaplastic oligodendroglioma; DA, diffuse astrocytoma; PC1, first principal component; PC2, second principal component.

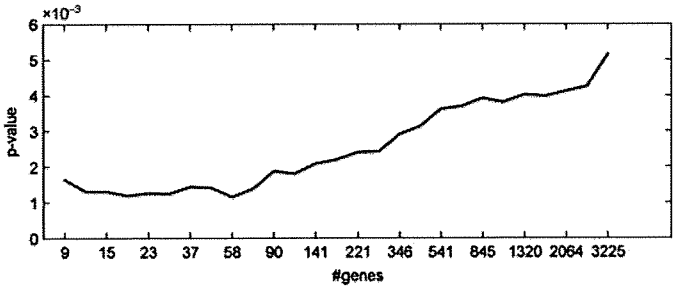

Fig. 2. Significance of the prediction model as a function of the number of diagnostic genes. The vertical axis is the *P*-value determined with the likelihood ratio test, whereas the horizontal axis is the number of top-ranked genes used for the model, selected using univariate Cox analysis.

relatively favorable GB cases with PFS periods of 2 years or more were located closer to the AO area (Fig. 1).

To further clarify the relationship between the gene expression profiles and the histology or prognosis of the tumors, we drew three axes (i.e. grade, oligo-astrocytic, and prognosis axes) in the scatter diagram in the two-dimensional principal component space (Fig. 1). The grade axis was determined by LDA, which discriminated grade 4 (GB) from grade 2 and 3 astrocytic tumors (DA and AA). The coordinate along the grade axis represented the correlation between each sample and its histological grade. The oligo axis was similarly determined by LDA, which discriminated between the oligodendroglial (AO) and astrocytic tumors (DA, AA, and GB). The prognosis axis was determined using multivariate Cox regression analysis. Because the direction of the prognosis axis was different from that of the grade axis, estimating the prognosis using gene expression patterns is likely to be better than estimations obtained using the histological grading.

We then constructed an outcome prediction model using a supervised method. Our prediction model was based on supervised PCA with the genes that were found to correlate with PFS using univariate Cox analysis. We evaluated the prediction model using a fivefold crossvalidation, in which the objective patients were left out of the supervised PCA process. We found that the best result was achieved when we used the 58 top-ranked genes (Fig. 2). The 58-gene model demonstrated a positive correlation between the PC1 score and PFS in the training set (Fig. 3a). When all of the training samples were divided into two groups based on the simple criterion that the coordinate on the PC1 score was positive or negative, a significant difference in PFS was observed between the two groups (Fig. 3b). We did not use any optimized thresholds either with the training or test sets. Even when the analysis was restricted to only the GB cases, PFS was significantly different between the two groups (Fig. 3c). Survival data for individual GB and non-GB patients, aligned based on the PC1 score, are shown in Figure 4a,b.

We selected 58 genes (listed in Table 1) using all of the training data, and constructed a prognosis predictor based on the PC1 score. This model successfully classified patients in the test set into good-prognosis and poor-prognosis groups (Fig. 3d,e,f). Survival data for each patient are shown in Figure 4c.

We examined the methylation status of the MGMT promoter in the GB patients. Among the 72 assessable cases (55 in the training set), the MGMT promoter was methylated in 40 of the cases (32 in the training set) (56%), and the methylation status was not obviously correlated with our prognosis predictor (Fisher's exact test $P = 0.3$).

We evaluated the 58-gene profile with other prognostic factors. First, we carried out Cox analysis with the 58-gene profile and tumor grade (GB vs others). In the univariate Cox analysis for PFS, we obtained $P = 2.7e^{-15}$ and $P = 1.0e^{-8}$ for the 58-gene profile and tumor grade, respectively. Multivariate Cox analysis including the 58-gene profile and tumor grade gave $1.0e^{-8}$ for the 58-gene profile, and 0.84 for tumor grade. In the univariate Cox analysis for OS, we obtained $6.0e^{-15}$ and $4.7e^{-9}$, for the 58-gene profile and tumor grade, respectively. Multivariate Cox analysis gave $2.4e^{-6}$ for the 58-gene profile, and 0.03 for tumor grade. Thus, the 58-gene profile is a strong prognostic factor independent of tumor grade.

We then carried out univariate and multivariate Cox regression analyses to evaluate the clinical parameters as potential predictors of PFS and OS among GB patients (Table 2). Univariate analysis revealed that the extent of resection, age, and the 58-gene profile were significantly correlated with OS, whereas only the 58-gene profile was significantly correlated with PFS. Multivariate analysis using the three factors extent of resection, 58-gene profile, and age showed that the extent of resection ($P = 0.0011$) and the 58-gene profile ($P = 0.0012$) were prognostic factors of similar strength (hazard ratio 3.1) for OS.

We checked the performance of our predictor using two publicly availably data sets: the MGH data set[12] and the MDA data set.[13] The correlation of the PC1 score to OS was evaluated with Cox regression tests, resulting in *P*-values of 0.00051 and 0.0066 for the MDA and MGH data sets, respectively. Our predictor produced a stable performance without over-fitting our data set. The results of Kaplan–Meier analysis are supplied as a supporting figure.

Among the 58 selected predictor genes, the expression of 37 genes was upregulated in the poor-prognosis group. They included *IGFBP2*, *VEGF*, *TNC*, *FN14*, *TIMP1*, *HMOX1*, *LGALS1*, and *UPAR*, all of which are known to be involved in angiogenesis or tumor-invasion processes. The remaining 21 genes showed

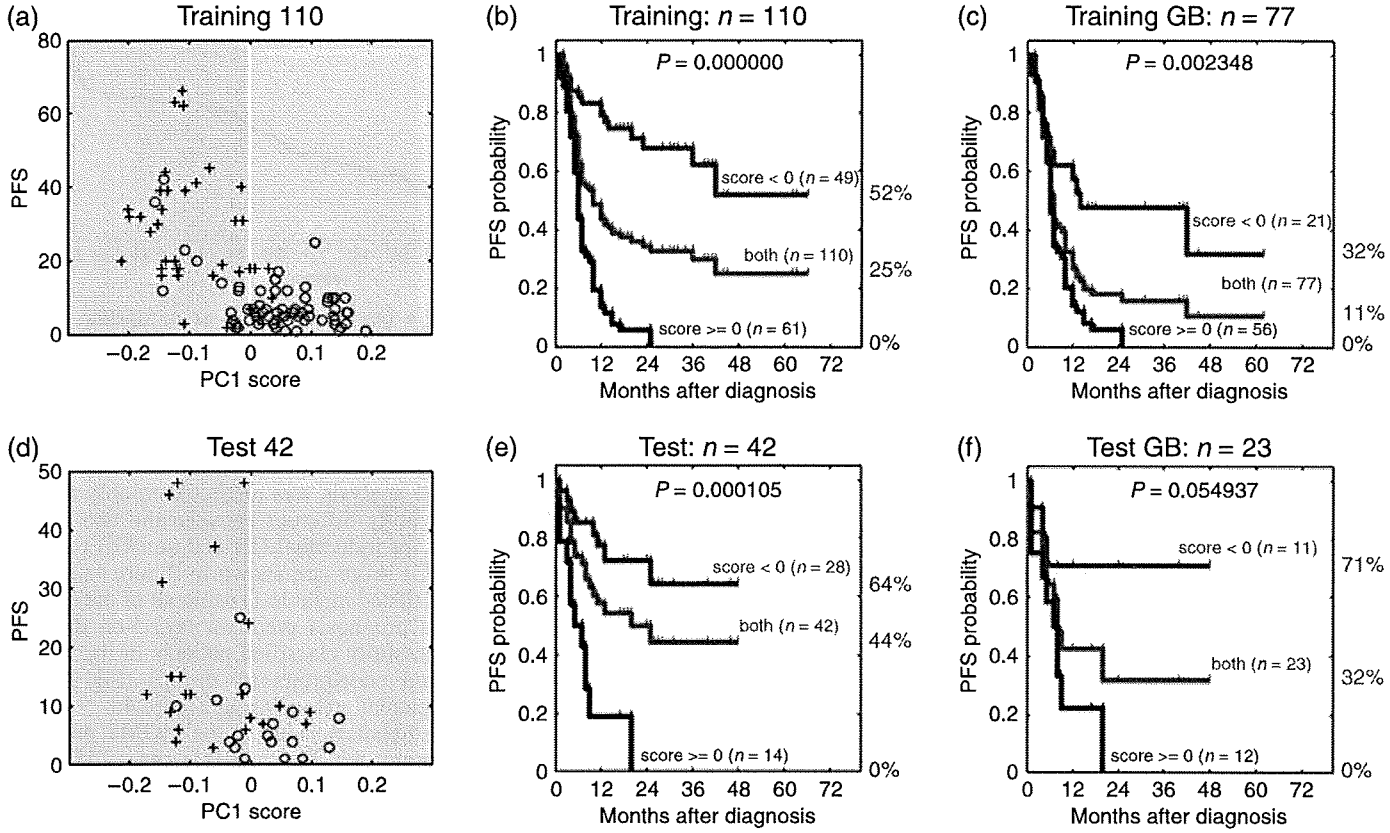

Fig. 3. (a,d) Correlations between the first principal component (PC1) score and progression-free survival (PFS) from (a) 110 gliomas in the training set, and from (d) 77 gliomas in the test set. Red circles and blue crosses denote the patients with tumor progression and the censored patients without tumor progression, respectively. (b,c,e,f) Kaplan–Meier estimates of PFS ratios: (b) 110 gliomas or (c) 77 glioblastomas (GB) in the training set; (e) 42 gliomas or (f) 23 GB in the test set.

upregulated expression levels in the good-prognosis group. Six of these genes (*INA*, *HES6*, *RTN1*, *BRSK2*, *SYN1*, and *CPLX2*) have been implicated in neuron-related functions (Table 1).

Discussion

Molecular-based classification of high-grade gliomas is expected to play an important role in predicting tumor prognosis, but it has been difficult to achieve practical applicability. Two gene expression profiling studies[13,22] constructed molecular classification schemes that correlated with survival. However, due to the lack of a proper comparison with other major prognostic factors,[23] the clinical utility of these schemes is unclear. In the present study, we enrolled patients mainly from a phase II clinical trial.[5] The treatment and collection of clinical information was carried out under strict guidelines, including frequent follow up and centralized diagnosis of MRI films. Using gene expression data obtained from high-throughput RT-PCR, we constructed a prognosis predictor that is independent of the primary prognostic factors. This prognosis predictor, composed of a 58-gene profile, was effective both for GB and non-GB cases. The system was a better predictor of PFS than OS, probably because PFS more directly correlates with the biological properties of gliomas.

Although our predictor was mainly based on the cases from the KNOG study, the results with two external data sets support the universal performance of the predictor irrespective of chemotherapeutic regimen. Because survival benefit by chemotherapy was relatively small in most malignant gliomas,[19] it is important to elucidate the differences in the intrinsic biological characters of the tumors.

Because the diagnostic genes were selected for correlation with prognosis, the prognostic predictor can be applied to any glioma, irrespective of histological grade. In our 52 non-GB cases, seven patients showed early progression within 6 months. By means of our prediction scheme, six out of the seven cases were classified into the poor-prognosis group, indicating good prognostic predictability for non-GB cases as well as GB cases.

Feature extraction by PCA and other techniques uncovered various molecular properties of the gliomas. AO localized to a particular area of the two-dimensional principal component space, indicating a distinct difference in the gene expression profiles of oligodendroglial and astrocytic tumors. The direction of the prognosis axis and the oligo axis indicated that different gene sets contributed to the differences in malignancy and histology. Differences in the directions of the prognosis and grade axes indicated that the pathological grading was not necessarily parallel to the refractoriness of the gliomas. Because the favorable GB cases were located closer to the AO cases, we speculate that deviation of the prognosis axis from the grade axis was due to the AO-like gene expression signature. This also agrees with our previous study on AO and GB classification:[4] 46 out of 168 diagnostic genes of the AO and GB classifier appeared among the 58 genes of the prognostic predictor. Those genes were simply selected by *P*-value-like scores for differential gene expression between AO and GB.

In our analyses, the expression levels of genes related to angiogenesis or invasion processes were higher in the poor-prognosis group. The expression of these genes has been reported to correlate with the malignant characteristics of gliomas, and some of them may work cooperatively.[24–33] Among the poor-prognosis markers reported by Phillips *et al.*,[13] *VEGF*,

doi: 10.1111/j.1349-7006.2008.01002.x

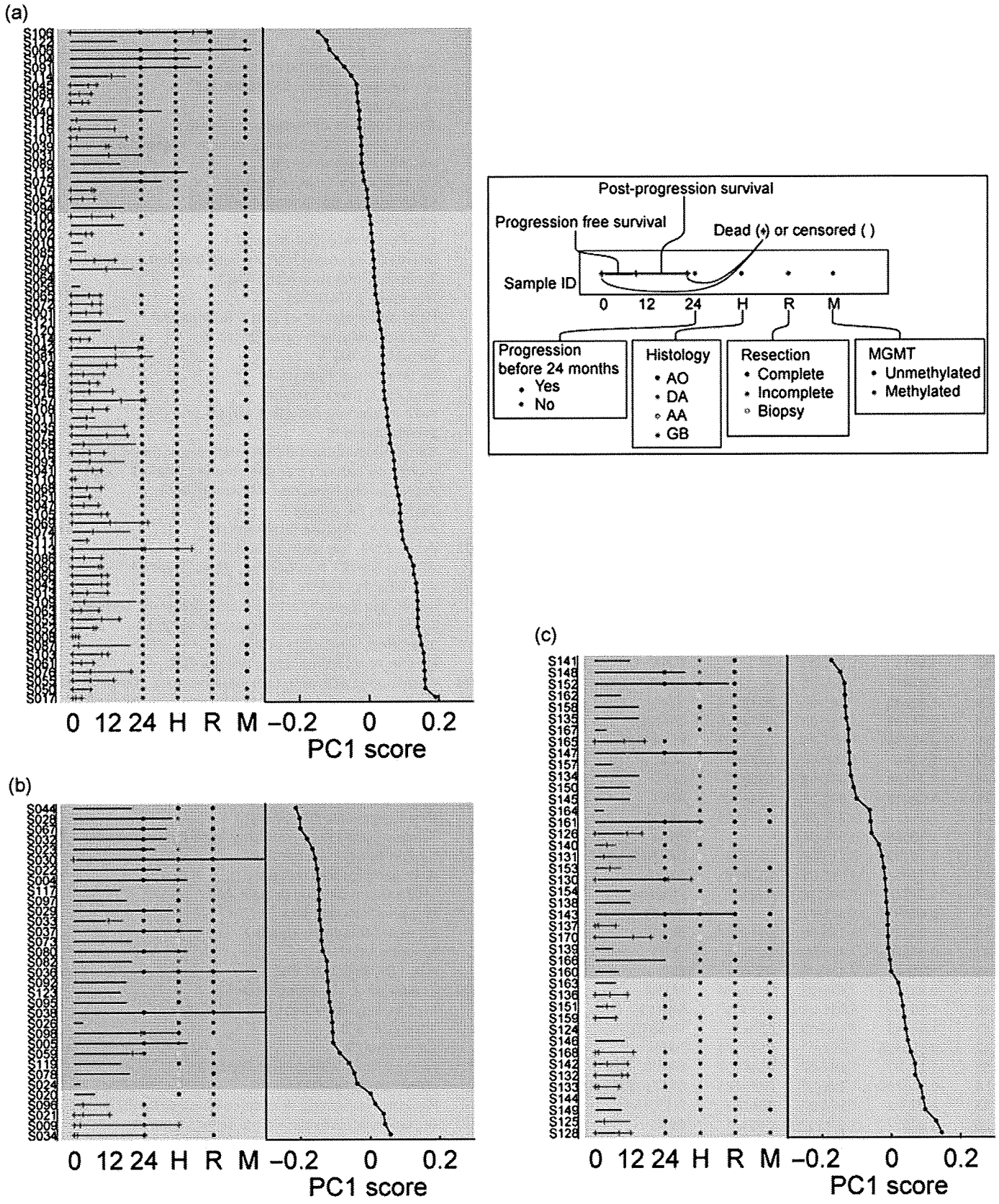

Fig. 4. Graphical representation of survival data for all of the patients sorted by the first principle component score. Explanatory notes are boxed. (a) Glioblastoma (GB) patients in the training set. (b) Non-GB patients in the training set. (c) Patients in the test set. AA, anaplastic astrocytoma; AO, anaplastic oligodendroglioma; DA, diffuse astrocytoma; MGMT, O^6-methyl guanine methyltransferase; PC1, first principle component.

Table 1. List of the 58 genes

No.	Coxβ	CoxP	GS number	Gene symbol	RefSeq ID	Description
1	1.58897	5.99E-14	GS2482	IGFBP2	NM_000597	Insulin-like growth factor binding protein 2
2	1.65957	6.14E-13	GS3909	VMP1	NM_030938	Hypothetical protein DKFZp566l133
3	1.99214	7.54E-13	GS10556	MSN	NM_002444	Moesin (MSN)
4	1.22527	9.07E-12	GS4923	TIMP1	NM_003254	Tissue inhibitor of metalloproteinase 1
5	1.78505	1.33E-11	GS1890	LGALS1	NM_002305	Lectin, galactoside-binding, soluble, 1 (galectin 1)
6	−1.39052	1.99E-11	GS12839	HMP19	NM_015980	HMP19 protein (HMP19)
7	−1.32313	9.28E-11	GS6687	CHGB	NM_001819	Chromogranin B (secretogranin 1)
8	2.61673	1.24E-10	GS3240	CD63	NM_001780	CD63 antigen (melanoma 1 antigen)
9	2.15797	2.30E-10	GS13698	NES	NM_006617	Nestin (NES)
10	1.62126	3.08E-10	GS2782	CLIC1	NM_001288	Chloride intracellular channel 1
11	−0.953909	3.10E-10	GS14040	INA	NM_032727	Internexin neuronal intermediate filament protein
12	1.23347	5.97E-10	–	TNC	NM_002160	Tenascin C (hexabrachion) (TNC)
13	1.48203	6.01E-10	GS4168	TAGLN2	NM_003564	Transgelin 2 (TAGLN2)
14	−1.20606	7.09E-10	GS13019	HES6	NM_018645	Hairy and enhancer of split 6 (*Drosophila*)
15	1.03772	1.64E-09	–	VEGF	BC065522	Vascular endothelial growth factor
16	1.24957	1.86E-09	–	VIM	NM_003380	Vimentin (VIM)
17	1.82864	2.30E-09	GS475	LDHA	NM_005566	Lactate dehydrogenase A
18	1.13676	2.47E-09	GS4232	TNC	NM_002160	Tenascin C (hexabrachion)
19	−1.18854	2.58E-09	GS13275	RPIP8	AB209802	RaP2 interacting protein 8 variant protein
20	−1.0301	4.30E-09	GS12811	SCG3	NM_013243	Secretogranin III
21	−1.05624	4.58E-09	GS14085	GDAP1L1	NM_024034	Ganglioside-induced differentiation-associated protein 1-like 1
22	1.48913	7.70E-09	GS1683	IFITM3	NM_021034	Interferon induced transmembrane protein 3 (1-8U)
23	2.98416	7.79E-09	GS1949	PPIB	NM_000942	Peptidylprolyl isomerase B (cyclophilin B)
24	2.53115	8.59E-09	GS421	TMSB4X	NM_021109	Thymosin, β4, X-linked
25	−1.09751	9.31E-09	GS10002	ALDOC	NM_005165	Aldolase C, fructose-bisphosphate
26	2.36402	9.99E-09	GS3483	ZYX	NM_001010972	Zyxin (ZYX), transcript variant 2
27	−1.00798	1.08E-08	GS13065	ATP1A3	NM_152296	ATPase, Na+/K+ transporting, α3 polypeptide
28	−1.09385	1.15E-08	–	ABCC8	NM_000352	ATP-binding cassette, sub-family C
29	0.982952	1.17E-08	GS6094	IGFBP3	NM_000598	Insulin-like growth factor binding protein 3, transcript variant 2
30	−1.33159	1.18E-08	GS13989	TUB	NM_003320	Tubby homolog (mouse), transcript variant 1
31	1.2851	1.30E-08	GS208	IFI30	NM_006332	Interferon, γ-inducible protein 30
32	1.5316	1.35E-08	–	FLNA	NM_001456	Filamin A, α (actin binding protein 280)
33	1.02419	1.52E-08	–	UPAR	NM_001005376	Plasminogen activator, urokinase receptor, transcript variant 2
34	1.50835	1.58E-08	GS13503	UPP1	NM_181597	Uridine phosphorylase 1 (UPP1), transcript variant 2
35	1.71108	1.59E-08	GS12786	LAMB2	NM_002292	Laminin, β2 (laminin S)
36	−0.989343	1.67E-08	GS13762	KIAA0927	AB023144	KIAA0927 protein
37	1.3664	1.68E-08	GS3760	AEBP1	NM_001129	AE binding protein 1
38	1.74194	1.84E-08	GS2836	EST	AJ420423	Full-length insert cDNA clone EUROIMAGE 1287006.
39	−1.00213	2.00E-08	GS14024	RTN1	NM_206857	Reticulon 1, transcript variant 2
40	1.1111	3.23E-08	GS11665	HMOX1	NM_002133	Heme oxygenase (decycling) 1
41	1.18646	3.34E-08	–	FN14	NM_016639	Tumor necrosis factor receptor superfamily, member 12A
42	−1.10155	3.71E-08	GS7227	DKFZp434J212	BC078676	Kinesin family member 21B
43	1.76858	4.94E-08	GS2958	GM2A	NM_000405	GM2 ganglioside activator
44	1.34778	5.34E-08	GS242	S100A10	NM_002966	S100 calcium binding protein A10
45	−0.824638	5.66E-08	–	PDE8B	AB085826	Phosphodiesterase 8B3
46	−1.05329	5.78E-08	GS13667	BRSK2	NM_003957	BR serine-threonine kinase 2
47	−0.956703	7.13E-08	GS4155	SYN1	M58378	Synapsin I (SYN1)
48	1.2386	7.13E-08	GS1071	EST	BX647603	cDNA DKFZp686L01105
49	−0.923475	7.67E-08	GS12884	CPLX2	NM_001008220	Complexin 2, transcript variant 2
50	1.50109	8.98E-08	GS1458	MRCL3	NM_006471	Myosin regulatory light chain MRCL3
51	1.92343	9.65E-08	GS2257	TMSB10	NM_021103	Thymosin, β10
52	−0.930736	1.20E-07	GS13880	JPH4	NM_032452	Junctophilin 4
53	−1.11747	1.26E-07	GS14607	FAIM2	NM_012306	Fas apoptotic inhibitory molecule 2
54	−0.909196	1.39E-07	GS11781	DKFZp761P2314	AL834342	cDNA DKFZp761P2314
55	1.23982	1.55E-07	GS6132	PLEKHA4	NM_020904	Pleckstrin homology domain containing, family A member 4
56	2.37577	1.67E-07	GS4131	GPX1	NM_000581	Glutathione peroxidase 1 transcript variant 1
57	1.15188	1.68E-07	GS2223	SOD2	NM_001024466	Superoxide dismutase 2, mitochondrial, transcript variant 3
58	1.93896	1.80E-07	GS7306	RHOC	NM_175744	*Homo sapiens* ras homolog gene family, member C

Coxβ, regression coefficient; CoxP, *P*-value for univariate Cox analysis.

VIM, and *NES* were included in our predictors. Another study also indicated that angiogenic activity represented by coexpression of *VEGF* and *IGFBP2* distinguished primary GB from secondary GB.[26] These genes and their protein products could be potential therapeutic targets in patients from the poor-prognosis group.

The good-prognosis group was characterized by upregulation of the expression of neuron-related genes. Phillips *et al.* also

doi: 10.1111/j.1349-7006.2008.01002.x

Table 2. Univariate and multivariate analyses for overall survival (OS) and progression-free survival (PFS) in glioblastoma patients

Variable	No. patients	OS		PFS	
		Hazard ratio (95% CI)	*P*-value	Hazard ratio (95% CI)	*P*-value
Univariate analysis					
Age <50 years versus ≥50 years	98	1.9 (1.1–3.6)	0.025	1.2 (0.71–2.1)	0.47
Extent of resection†	98	3.1 (1.5–6.7)	0.00085	1.4 (0.79–2.5)	0.23
MGMT‡	72	0.78 (0.43–1.40)	0.43	0.67 (0.39–1.2)	0.16
58 gene profile	98	3.8 (1.8–.9)	0.000051	3.0 (1.7–5.6)	0.0001
Multivariate analysis					
Age <50 years versus ≥50 years	98	1.8 (0.93–3.40)	0.078	–	–
Extent of resection†	98	3.1 (1.5–6.7)	0.0011	–	–
58 gene profile	98	3.1 (1.5–6.5)	0.0012	–	–

†Partial resection and biopsy versus complete resection. ‡Methylated versus unmethylated O6-methyl guanine methyltransferase (MGMT) promotor. CI, confidence interval.

described a correlation between neuronal markers and the favorable subclasses.[13] Likewise, other investigators have reported that a subset of neuronal genes was highly expressed in AO tumors with better prognoses.[34,35] Taken together, these results indicate that the expression of neuron-related genes is a marker of good prognoses in patients with high-grade gliomas.

Among the 58 genes in the predictor model, *IGFBP2* and *VEGF* also appeared in the 44-gene classifier described by Freije *et al.*,[22] whereas *TIMP1* and *SCG3* appeared in the 35-gene signature developed by Phillips *et al.*[13] Although most of the genes did not overlap, these three gene sets might have similar prognostic value, because distinct but equally predictive gene lists can be derived from the same data matrix.[36] It should be noted, however, that the 35-gene signature was strongly correlated with age ($P < 0.005$), making its clinical utility uncertain.

Hegi *et al.* recently demonstrated that epigenetic silencing of the MGMT gene serves as an independent prognostic parameter in GB patients treated with temozolomide.[3] In our analysis, the methylation status of the MGMT promoter did not prove to be a significant prognostic factor. Kamiryo *et al.* found that MGMT methylation was a significant prognostic factor for both for OS and PFS in patients with grade III tumors, but not for grade IV tumors.[37] Our result is consistent with this report, suggesting that the effect of MGMT on ACNU-based treatment is likely to be smaller than that on temozolomide in GB patients.

Why is the extent of resection a significant prognostic factor for OS but not for PFS? Extent of resection has been reported as an independent prognostic factor to OS (e.g. Lacroix *et al.*[38]), but even a stringently designed study reported insignificance both for OS and PFS.[39] We carefully evaluated individual patients and found that complete resection cases were more accessible to second surgery and survival after progression was extended. Due to careful follow up, tumor progression was detected earlier than usual, and the number of complete resection cases with second operation was increased. This explains the discrepancy between OS and PFS.

Another concern is whether performance of the predictor for OS would be improved if OS data were used for training. We examined this possibility, but found that performance for PFS was consistently better than that for OS, irrespective of the training data. This also suggests a tight link between PFS and gene expression. The details of the analysis will be described elsewhere.

In conclusion, our profiling results will help to construct a new classification scheme that assesses clinical malignancies better than the conventional histological classification system.

Acknowledgments

We would like to thank Dr H. Kita-Matsuo for technical advice, and Mrs S. Maki-Kinjo, Mrs K. Miyaoka-Ikenaga and Mrs Y. Ishida for their valuable technical assistance. This work was supported by a Grant-in-Aid for the Development of Innovative Technology from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

References

1 Louis DN, Holland EC, Cairncross JG. Glioma classification: a molecular reappraisal. *Am J Pathol* 2001; **159**: 779–86.
2 Cairncross JG, Ueki K, Zlatescu MC *et al.* Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *J Natl Cancer Inst* 1998; **90**: 1473–9.
3 Hegi ME, Diserens AC, Gorlia T *et al.* MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* 2005; **352**: 997–1003.
4 Shirahata M, Iwao-Koizumi K, Saito S *et al.* A gene expression-based molecular diagnostic system for malignant gliomas displays clinical utility, prognostic ability and reproducibility superior to histological diagnosis. *Clin Cancer Res* 2007; **13**: 7341–56.
5 Aoki T, Takahashi JA, Ueba T *et al.* Phase II study of nimustine, carboplatin, vincristine, and interferon-β with radiotherapy for glioblastoma multiforme: experience of the Kyoto Neuro-Oncology Group. *J Neurosurg* 2006; **105**: 385–91.
6 Kleihues P, Cavenee WK. *World Health Organization Classification of Tumours: Pathology and Genetics: Tumours of the Nervous System*. Lyon: IARC Press, 2000.
7 Matoba R, Kato K, Saito S *et al.* Gene expression in mouse cerebellum during its development. *Gene* 2000; **241**: 125–31.
8 Iwao-Koizumi K, Matoba R, Ueno N *et al.* Prediction of docetaxel response in human breast cancer by gene expression profiling. *J Clin Oncol* 2005; **23**: 422–31.
9 Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci USA* 1996; **93**: 9821–6.
10 Japan CoBTRo. Report of Brain Tumor Registry of Japan (1969–96). *Neurologia Medico-Chirurgica* 2003; **43** Suppl: 1–111.
11 Oba S, Sato MA, Takemasa I, Monden M, Matsubara K, Ishii S. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics (Oxford, England)* 2003; **19**: 2088–96.
12 Nutt CL, Mani DR, Betensky RA *et al.* Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* 2003; **63**: 1602–7.
13 Phillips HS, Kharbanda S, Chen R *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* 2006; **9**: 157–73.
14 Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci USA* 2000; **97**: 10 101–6.
15 Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *Plos Biol* 2004; **2**: E108.

16 Medical Research Council Brain Tumour Working Party. Randomized trial of procarbazine, lomustine, and vincristine in the adjuvant treatment of high-grade astrocytoma: a Medical Research Council trial. *J Clin Oncol* 2001; **19**: 509–18.
17 Chang CH, Horton J, Schoenfeld D *et al.* Comparison of postoperative radiotherapy and combined postoperative radiotherapy and chemotherapy in the multidisciplinary management of malignant gliomas. A joint Radiation Therapy Oncology Group and Eastern Cooperative Oncology Group study. *Cancer* 1983; **52**: 997–1007.
18 Fine HA, Dear KB, Loeffler JS, Black PM, Canellos GP. Meta-analysis of radiation therapy with and without adjuvant chemotherapy for malignant gliomas in adults. *Cancer* 1993; **71**: 2585–97.
19 Stewart LA. Chemotherapy in adult high-grade glioma: a systematic review and meta-analysis of individual patient data from 12 randomised trials. *Lancet* 2002; **359**: 1011–18.
20 Stupp R, Mason WP, van den Bent MJ *et al.* Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med* 2005; **352**: 987–96.
21 Walker MD, Green SB, Byar DP *et al.* Randomized comparisons of radiotherapy and nitrosoureas for the treatment of malignant glioma after surgery. *N Engl J Med* 1980; **303**: 1323–9.
22 Freije WA, Castro-Vargas FE, Fang Z *et al.* Gene expression profiling of gliomas strongly predicts survival. *Cancer Res* 2004; **64**: 6503–10.
23 Curran WJ Jr, Scott CB, Horton J *et al.* Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *J Natl Cancer Inst* 1993; **85**: 704–10.
24 Bussolati B, Ahmed A, Pemberton H *et al.* Bifunctional role for VEGF-induced heme oxygenase-1 *in vivo*: induction of angiogenesis and inhibition of leukocytic infiltration. *Blood* 2004; **103**: 761–6.
25 Deryugina EI, Bourdon MA. Tenascin mediates human glioma cell migration and modulates cell migration on fibronectin. *J Cell Sci* 1996; **109 (Pt 3)**: 643–52.
26 Godard S, Getz G, Delorenzi M *et al.* Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Cancer Res* 2003; **63**: 6613–25.
27 Gondi CS, Lakka SS, Yanamandra N *et al.* Expression of antisense uPAR and antisense uPA from a bicistronic adenoviral construct inhibits glioma cell invasion, tumor growth, and angiogenesis. *Oncogene* 2003; **22**: 5967–75.
28 Nishie A, Ono M, Shono T *et al.* Macrophage infiltration and heme oxygenase-1 expression correlate with angiogenesis in human gliomas. *Clin Cancer Res* 1999; **5**: 1107–13.
29 Plate KH, Breier G, Weich HA, Risau W. Vascular endothelial growth factor is a potential tumour angiogenesis factor in human gliomas *in vivo*. *Nature* 1992; **359**: 845–8.
30 Rorive S, Belot N, Decaestecker C *et al.* Galectin-1 is highly expressed in human gliomas with relevance for modulation of invasion of tumor astrocytes into the brain parenchyma. *Glia* 2001; **33**: 241–55.
31 Song SW, Fuller GN, Khan A *et al.* IIp45, an insulin-like growth factor binding protein 2 (IGFBP-2) binding protein, antagonizes IGFBP-2 stimulation of glioma cell invasion. *Proc Natl Acad Sci USA* 2003; **100**: 13 970–5.
32 Tran NL, McDonough WS, Savitch BA *et al.* Increased fibroblast growth factor-inducible 14 expression levels promote glioma cell invasion via rac1 and nuclear factor-κB and correlate with poor patient outcome. *Cancer Res* 2006; **66**: 9535–42.
33 Wang H, Wang H, Shen W *et al.* Insulin-like growth factor binding protein 2 enhances glioblastoma invasion by activating invasion-enhancing genes. *Cancer Res* 2003; **63**: 4315–21.
34 Mukasa A, Ueki K, Ge X *et al.* Selective expression of a subset of neuronal genes in oligodendroglioma with chromosome 1p loss. *Brain Pathol (Zurich, Switzerland)* 2004; **14**: 34–42.
35 Mukasa A, Ueki K, Matsumoto S *et al.* Distinction in gene expression profiles of oligodendrogliomas with and without allelic loss of 1p. *Oncogene* 2002; **21**: 3961–8.
36 Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics (Oxford, England)* 2005; **21**: 171–8.
37 Kamiryo T, Tada K, Shiraishi S, Shinojima N, Kochi M, Ushio Y. Correlation between promoter hypermethylation of the O6-methylguanine-deoxyribonucleic acid methyltransferase gene and prognosis in patients with high-grade astrocytic tumors treated with surgery, radiotherapy, and 1-(4-amino-2-methyl-5-pyrimidinyl) methyl-3-(2-chloroethyl)-3-nitrosourea-based chemotherapy. *Neurosurgery* 2004; **54**: 349–57.
38 Lacroix M, Abi-Said D, Fourney DR *et al.* A multivariate analysis of 416 patients with glioblastoma multiforme: prognosis, extent of resection, and survival. *J Neurosurgery* 2001; **95**: 190–8.
39 Athanassiou H, Synodinou M, Maragoudakis E *et al.* Randomized phase II study of temozolomide and radiotherapy compared with radiotherapy alone in newly diagnosed glioblastoma multiforme. *J Clin Oncol* 2005; **23**: 2372–7.

Supporting information

Additional supporting information may be found in the online version of this article:

Fig. S1. Kaplan–Meier analysis of publicly available data sets. Online only.

Table S1. Gene expression data of the 3456 genes. Annotation information is also included. Online only.

Table S2. Clinical information about the patients in the training set. Online only.

Table S3. Clinical information about the patients in the test set. Online only.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

doi: 10.1111/j.1349-7006.2008.01002.x

Optimal Aggregation of Binary Classifiers for Multiclass Cancer Diagnosis Using Gene Expression Profiles

Naoto Yukinawa, Shigeyuki Oba, Kikuya Kato, and Shin Ishii

Abstract—Multiclass classification is one of the fundamental tasks in bioinformatics and typically arises in cancer diagnosis studies by gene expression profiling. There have been many studies of aggregating binary classifiers to construct a multiclass classifier based on one-versus-the-rest (1R), one-versus-one (11), or other coding strategies, as well as some comparison studies between them. However, the studies found that the best coding depends on each situation. Therefore, a new problem, which we call the "optimal coding problem," has arisen: how can we determine which coding is the optimal one in each situation? To approach this optimal coding problem, we propose a novel framework for constructing a multiclass classifier, in which each binary classifier to be aggregated has a weight value to be optimally tuned based on the observed data. Although there is no a priori answer to the optimal coding problem, our weight tuning method can be a consistent answer to the problem. We apply this method to various classification problems including a synthesized data set and some cancer diagnosis data sets from gene expression profiling. The results demonstrate that, in most situations, our method can improve classification accuracy over simple voting heuristics and is better than or comparable to state-of-the-art multiclass predictors.

Index Terms—Multiclass classification, error correcting output coding, gene expression profiling, cancer diagnosis.

1 Introduction

DNA microarrays or alternative quantification techniques have enabled genome-wide expression analyses of various biological phenomena. One important application of this technique is cancer diagnosis, where the expression level of thousands of genes can be used as a vast amount of molecular biomarkers of specific phenotypes. This analysis is expected to overcome the conventional problems of histopathological cancer diagnosis such as variations in diagnosis by individual pathologists or difficulties in differentiating between malignant and benign tissues due to their morphological similarities. For constructing diagnosis systems using high-dimensional gene expression data, supervised learning theories are often applied, and several studies have been successful in recent years. Representative studies include classification of two kinds of acute leukemias [1] by weighted voting algorithm, classification of four types of small round blue cell tumors (SRBCTs) by artificial neural networks [2], and the diagnosis of multiple (14 types) common adult malignancies by a multiclass support vector machine (SVM) [3]. These existing studies revealed that tissues from different origins can be well classified by supervised classification algorithms, mainly because the gene expression profile of an origin is considerably different from the others. On the contrary, classifying multiple types of tissue from the same origin, for example, hereditary breast cancer [4], is much more difficult due to the similarity in gene expression patterns between phenotypic variants; there is still no definitive method. When considering histopathological applications in the postgenomic era, however, we must deal with such difficult situations, and sophisticated multiclass prediction methods are required. In this paper, we propose a novel supervised learning approach to multiclass classification problems.

For classifying gene expression profiles, SVM is thought to be the most promising method in recent years, because a larger margin of decision boundary between two classes improves its generalization capability for class separation, especially in a high-dimensional gene expression vector space. SVM can originally handle binary classification problems. In a multiclass problem, however, it needs some device to integrate the binary classification results into the final answer to the original multiclass (M classes) classification problem. For the integration process, the following simple voting heuristics have been frequently used: 1) prepare a set of M binary classifiers, each of which separates one class from the other classes (one-versus-the-rest: 1R); then, a single guess is determined by voting the outputs from the M binary classifiers [5] and 2) prepare a set of $M(M-1)/2$ binary classifiers, each of which separates one class from another (one-versus-one: 11); then, a single guess is determined by a vote performed by them [6]. These integration processes are generalized

- *N.Yukinawa and S. Oba are with the Graduate School of Information Sciences, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan. E-mail: {naoto-yu, shige-o}@is.naist.jp.*
- *K. Kikuya us wuth the Research Institute, Osaka Medical Center for Cancer and Cardiovascular Diseases, 1-3-2 Nakamichi, Higashinari-ku, Osaka 537-8511, Japan. E-mail: katou-k@mc.pref.osaka.jp.*
- *S. Ishii is with the Graduate School of Informatics, Kyoto University, Gokajo, Uji, Kyoto 611-0011, Japan, and the Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-chi, Ikoma, Nara 630-0192, Japan. E-mail: ishii@i.kyoto-u.ac.jp.*

Manuscript received 18 May 2006; revised 25 May 2007; accepted 29 June 2007; published online 31 July 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0113-0506.
Digital Object Identifier no. 10.1109/TCBB.2007.70239.

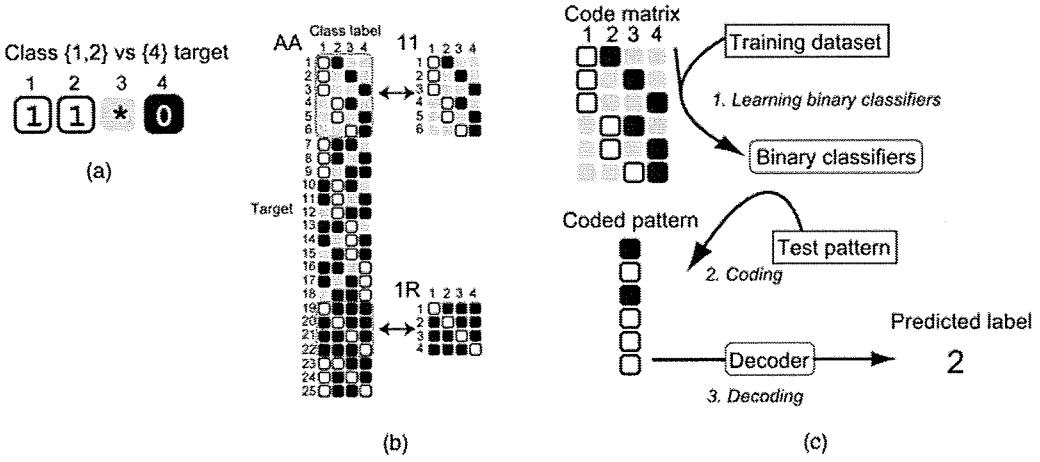

Fig. 1. Overview of ECOC method. (a) An example of "target" for a four class problem that represents the corresponding set of binary classification problems. A target is represented by a three-valued row vector of length 4 (which corresponds to the number of classes), where each column's index corresponds to a single class label. "1" (white square) and "0" (black square) indicate positive and negative class labels for binary classification (in this case, {1, 2} versus {4}), respectively, and "$*$" (gray square) indicates unused class labels. (b) Typical code matrices for the four class problem. A code matrix consists of arbitrary target vectors, i.e., the row and column indices correspond to a target and a class label, respectively. 1R and 11 code matrices are traditional designs; 1R is the set of one-versus-the-rest targets ({1} versus {1, 2, 3}, {2} versus {1, 3, 4}, $\cdots$), and 11 is the set of one-versus-one targets ({1} versus {2}, {1} versus {3}, $\cdots$). AA consists of all possible targets including 1R and 11. (c) Multiclass classification by ECOC method. In this example, the 11 code matrix and the Hamming decoder are used. First, six binary classifiers, each 11 target, are trained based on a training data set. Then, a test pattern is classified by the six classifiers, and consequently, the binary (coded) pattern is obtained. The decoder searches for the nearest column vector (code word) in the designed 11 code matrix with respect to Hamming distance and outputs the corresponding class label as the final guess.

down to the framework of error correcting output coding (ECOC) [7], which enables the use of a general set of binary classifiers such as exhaustive coding [7] and random coding [8]. In addition, arbitrary integration methods rather than simple voting can be implemented in the ECOC framework. For example, Hastie and Tibshirani [9] proposed a probabilistic approach, which made it possible to integrate probabilistic outputs from binary classifiers of 11 coding. Zadrozny [10] also presented a probabilistic approach to integrate a general set of binary outputs.

There are also some comparison studies of these various classification methods applied to multiclass cancer classification problems. Li et al. [11] compared the performance of several multiclass classification methods by applying them to published data sets of gene expression profiles; they evaluated SVMs including simple voting heuristics with 1R, 11, exhaustive, and random coding, as well as the Naive Bayes method, KNN, and the J4.8 decision tree. They found that SVMs showed overwhelming performance in most cases and that choosing a set of binary classifiers, i.e., favorable coding, was problem specific. Ramaswamy et al. [3] also compared the performance of SVMs with 1R and 11 and concluded that 1R showed better performance. Statnikov et al. [12] exhaustively compared the performance of several SVMs, KNN, and neural networks by using published gene expression data sets, concluding that multiclass SVMs [13], [14] and simple voting (1R) were the better classification methods; however, the best SVM algorithm among them was again problem specific.

In this study, we propose a novel framework to obtain problem-specific optimal coding. We first revisit the probabilistic approach proposed in [9], leading to our modification called the maximum a posteriori (MAP) method. In order to deal with the optimal coding problem, then, we introduce weights to the constituent binary classifiers, which are optimized so as to maximize the classification performance for the training data set; this is called a weighted MAP (WMAP) algorithm. It can obtain a better "graded" set of binary classifiers than the conventional 1R and 11 by solving the optimal coding problem. We show that the proposed method improves classification performance over simple voting heuristics by binary classifiers not only for a synthesized problem but also for several difficult multiclass cancer classification problems.

2 ECOC and Optimal Coding Problem

The primary objective of supervised multiclass prediction is to construct a predictor that predicts the class label $i^{(n)} \in C$ of the nth sample from its pattern vector $x^{(n)}$, where $C = \{1, \ldots, M\}$ is a set of $M \geq 3$ class labels. The predictor is constructed based on the training data set consisting of N samples accompanied by their class labels, $L = \{x^{(n)}, i^{(n)}\}_{n=1,\ldots,N}$.

In the framework of ECOC [7], [15], each multiclass problem is decomposed into multiple binary prediction problems, which are denoted by a code matrix $\{"1," "0," "*"\}^{l \times M}$, where l represents the number of binary prediction problems (see Fig. 1). We call the configuration of a code matrix "coding" or "coding method." For example (Fig. 1a), when the jth row of the code matrix includes "1" as the first and second elements, "0" as the fourth element, and "$*$" as the third element, this row indicates that the jth binary predictor ideally outputs "0" and "1" for input sample patterns belonging to classes {1, 2} and {4}, respectively; the jth predictor does not care about the sample patterns belonging to the third class. We call the pair of subsets corresponding to

the jth row of the code matrix, {1, 2} and {4} in this case, the jth "target." For an input sample x whose multiclass label is being predicted, multiple outputs from all the binary predictors defined by the code matrix are aggregated and "decoded" into a multiclass output i if the set of binary outputs was most similar to the ith column of the code matrix, which is termed "code word" (Fig. 1c). Although some binary predictors may make errors in actual cases, if the number of errors is not too large, an appropriate "decoding procedure" can correct the errors to restore the correct multiclass label. This is the basic idea of ECOC.

To design an effective classifier according to the ECOC framework, selecting the appropriate "code matrix" and "decoding procedure" is essential. Conventional procedures of multiclass prediction based on the one-versus-the-rest (**1R**) or one-versus-one (**11**) methods are understood as practical examples of ECOC employing simple code matrices representing 1R or 11 (Fig. 1b) and the simplest Hamming decoding procedure. In the ECOC framework, favorable coding can be selected from the heuristics candidates such as 11 and 1R and all-possible-combinations (**AA**, see Fig. 1b). Although an optimal coding (or an optimal code matrix), if it exists, is expected to enhance the resultant multiclass prediction, which code matrix is the optimal one has been found to depend on each situation [16]. In our study, instead of looking directly for the optimal coding, we intend to optimally weigh the binary classifiers whose set is given arbitrarily as an initial code matrix. Since this weight optimization is performed so as to exhibit the best performance based on a given data set, we expect that the optimal coding problem can be solved in a consistent manner in each situation. The validity of this novel idea is examined through experiments using a synthesized data set and some difficult bioinformatics data sets.

3 Combining Probabilistic Guesses of Binary Classifiers by Statistical Estimation

Our framework employs a probabilistic decoding, which was first proposed by Hastie and Tibshirani [9], in particular, for 11 coding and later extended by Zadrozny [10] as a general coding method. It decodes a probabilistic guess on the multiclass problem from the aggregated probabilistic guesses on the binary problems.

For the nth sample with a sample pattern vector $x^{(n)}$, we assume a class membership probability vector $p^{(n)}$ whose component is a true but unobserved membership probability $p_i^{(n)}$ to each class label $i \in C$:

$$p_i^{(n)} \geq 0, \quad \sum_{i \in C} p_i^{(n)} = 1. \tag{1}$$

We attempt to estimate $p^{(n)}$ and call the estimate a probabilistic guess of the primary multiclass problem. Let $q_j^{(n)} \equiv Pr(i \in 1_j | x^{(n)}, i \in 1_j \cup 0_j)$ be a probabilistic guess of the jth binary predictor to the nth sample, where $1_j \subset C$ and $0_j \subset C$ are class subsets corresponding to the binary outputs "1" (positive) and "0" (negative) of the jth binary predictor, respectively.[1] Let $q^{(n)} = \{q_j^{(n)}\}_{j \in B}$ denote the set of class membership probabilities, where B is the set of binary predictors defined by a code matrix. It is noted that the code matrix B can be represented by an arbitrary set of code words (each of which corresponds to a class), not restricted as 1R or 11, according to our approach (Fig. 1b). Thus, the class membership probability vector for the entire data set, $\{q^{(n)}\}_{n=1,\cdots,N}$, is determined by a set of binary classifiers in B, based on the training data set L. In the following, we omit the argument "$^{(n)}$" when that does not risk causing confusion.

Since our study aims at presenting a good methodology to deal with the optimal coding problem, our task is, in principle, free from the choice of binary classifiers. For frequently used binary classifiers such as linear discriminant analysis and SVM, probabilistic outputs are not available straightforwardly. In this study, we use SVM as an individual binary classifier, to which we apply logistic regression whose parameter is determined by cross validation with the training data set [17], in order to obtain a probabilistic guess from the discriminant function value of the SVM (for details, see Appendix A). The dependence on individual binary classifiers will be briefly discussed in Section 6.

Next, we proceed to an estimation procedure of multiclass membership p from the set of binary membership probabilities q. Based on the assumption of the true multiclass membership probability p, the true binary class probability with respect to the jth target, $\pi_j(x) = Pr(i \in 1_j | x, i \in 1_j \cup 0_j)$, is given by

$$\pi_j = \frac{p_{1_j}}{p_{1_j} + p_{0_j}}, \tag{2}$$

where membership probability, p_l, to a subset of class labels $l \in \tilde{2}^C$ is given by a simple summation of class membership probabilities to single classes, $p_l = \sum_{i \in l} p_i$. To obtain a p, which allows π to best fit the observed q, a weighted Kullback-Leibler (KL) divergence between q and π is minimized with respect to p:

$$KL(q; \pi(p)) = \sum_{j \in B} w_j \left\{ q_j \log \frac{q_j}{\pi_j} + (1 - q_j) \log \frac{1 - q_j}{1 - \pi_j} \right\}, \tag{3}$$

where w_j is a confidence weight variable corresponding to the jth target, which could be set at $w_j = 1$ in the simplest case. In the next section, we will consider how to determine the w_j value appropriately, which corresponds to the optimal coding process. Since the natural distribution of p is multinomial, we introduce a Dirichlet prior to (3) for regularization, and the problem is formulated as maximization of the following objective function:

$$V(p) = \sum_{j \in B} w_j \{ q_j \log p_{1_j} + (1 - q_j) \log p_{0_j} - \log(p_{1_j} + p_{0_j}) \} + \sum_{i \in C} \gamma_0 \log p_i + R, \tag{4}$$

where γ_0 is a hyperparameter that controls the intensity of the Dirichlet prior, and R is a constant independent of p. The Dirichlet prior term controls prior knowledge of the rate of random mislabels and contributes to stabilizing the

1. In Fig. 1a, $1_j = \{1, 2\}$ and $0_j = \{4\}$. In Fig. 1b, 1_j and 0_j are denoted by white and black squares, respectively.

optimization algorithm. We set $\gamma_0 = 0.001$ in this study, which leads to stability, whereas its variation did not affect the results very much. By maximizing objective function $V(\boldsymbol{p})$ with respect to $\boldsymbol{p}$ under constraint (1), we obtain the probability estimate of class membership $\hat{\boldsymbol{p}}$. This maximization can be performed by the steepest descent method with a Lagrange multiplier. In the simplest case where all the weight variables $\{w_j\}_{j\in B}$ are set at unity, this probabilistic estimation is similar to the existing probabilistic decoding [9] and is subsequently called the MAP method. The pseudocode for the MAP method is presented as Algorithm 1 in Appendix B.

4 Optimization of the Weights of Binary Classifiers

In this section, we propose a procedure to optimize the weight variable $\boldsymbol{w} = \{w_j\}_{j\in B}$, which allows us to approach the optimal coding within the usage of initial code matrix B. To optimize the weight $\boldsymbol{w}$, we define a gain function U that represents the concordance between the class membership probability estimate $\boldsymbol{p}$ and the true class label i:

$$U \equiv U(\{\boldsymbol{p}^{(n)}\}_{n=1,\cdots,N}, \{t^{(n)}\}_{n=1,\ldots,N}) = \sum_{n=1}^{N}\sum_{i\in C} t_i^{(n)} \mathrm{mx}(p_i^{(n)}), \tag{5}$$

where $t^{(n)} = (t_1^{(n)}, \ldots, t_M^{(n)})$ is an M-dimensional binary vector that indicates a single class label; $t_i^{(n)} = 1$ if sample n belongs to class i, otherwise, $t_i^{(n)} = 0$. $\mathrm{mx}(\boldsymbol{p}_i)$ is a soft-max function:

$$\mathrm{mx}(p_i) = \frac{\exp(\beta p_i)}{Z}, \quad Z = \sum_{i'\in C} \exp(\beta p_{i'}),$$

where β is an inverse temperature parameter, which controls the sharpness of the soft-max function; as $\beta \to +\infty$, $\mathrm{mx}(p_i)$ approaches 1 for $i = \arg\max_i p_i$, or 0 otherwise. Since the setting of this parameter barely affects the results, we set it at an appropriately large value.

The MAP solution does not depend on any linear scale of the KL divergence (3), i.e., multiplication of every weight w_j, $j \in B$ by a constant. To remove this scale insensitivity, we introduce a constraint:

$$w_j \geq 0, \quad \sum_{j\in B} w_j = 1. \tag{6}$$

The gain function U is an implicit function of $\boldsymbol{w}$, namely, U depends on $\boldsymbol{p}$, which is obtained by maximizing a function of $\boldsymbol{w}$. Therefore, the optimization of U with respect to $\boldsymbol{w}$ is to obtain the $\tilde{\boldsymbol{w}}$ that satisfies

$$\tilde{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} U(\{\tilde{\boldsymbol{p}}(\boldsymbol{w})^{(n)}\}_{n=1,\ldots,N}, \{t^{(n)}\}_{n=1,\ldots,N}) \quad \text{under condition (6)}, \tag{7}$$

$$\tilde{\boldsymbol{p}}^{(n)} = \arg\max_{\boldsymbol{p}^{(n)}} V(\boldsymbol{p}^{(n)}|\boldsymbol{w}) \text{ under condition (1)}, \tag{8}$$

for the given training data set $L \equiv \{\boldsymbol{q}^{(n)}, t^{(n)}\}_{n=1,\ldots,N}$. This is a twofold optimization problem; outer and inner optimization is given by (7) and (8). The optimal M-class classifier is configured by optimizing $\tilde{\boldsymbol{w}}$ for the entire data set L in the outer optimization, and by using it, the class membership probability estimate $\boldsymbol{p}^{(n)}$ of each pattern vector $\boldsymbol{x}^{(n)}$ is given in the inner optimization. In other words, the outer and inner optimization corresponds to the optimal coding and the decoding processes, respectively.

A solution to this optimization problem is shown in Appendix C. We call this algorithm the WMAP method. Note that we can utilize an arbitrary gain function in place of (5), if the gain function is differentiable with respect to $\boldsymbol{p}^{(n)}$. Accordingly, our WMAP approach looks for the optimal "graded" coding represented as the weight vector $\boldsymbol{w}$ within the initial setting of the "binary" code matrix B. This optimization process is in principle free from the choice of individual binary classifiers (typically SVMs), probabilistic transformation from their discriminant function (typically logistic regression), and the original code matrix (typically AA). The pseudocode for this weight optimization procedure is presented as Algorithm 2 in Appendix B.

5 Results

5.1 Experiment 1: Applications to Synthesized Data Sets

We first examined the performance of the two methods, MAP and WMAP, by applying them to a synthesized data set. The aim of this experiment is to show performance improvement by WMAP in each of three designs of code matrix: 1R, 11, or AA. Assuming an underlying 3-class structure of 2D data points, the data set was synthesized according to the following procedure. First, we generated each data point $\boldsymbol{x} = (x_1, x_2)$ from a 2D uniform distribution within $[-2, 2] \times [-2, 2]$. Next, the class label of each data point was assigned as $\arg\min \|\boldsymbol{x} - \boldsymbol{x}_{c_i}\|^2 - b_{c_i}$ based on the distance between the data point and the centroids of the three classes: $\boldsymbol{x}_{c_1} = (-\sqrt{2}, -\sqrt{2})$, $\boldsymbol{x}_{c_1'} = (\sqrt{2}, \sqrt{2})$, $\boldsymbol{x}_{c_2} = (-\sqrt{2}, \sqrt{2})$, and $\boldsymbol{x}_{c_3} = (\sqrt{2}, -\sqrt{2})\}$, where $b_{c_1} = 2\log(0.35)$, $b_{c_1'} = 2\log(0.20)$, $b_{c_2} = 2\log(0.50)$, and $b_{c_3} = 2\log(0.75)$. Note that c_1 and c_1' represent the same class: that is, this class has two class centroids. We generated 400 points (c_1, c_1', c_2, and c_3: 100 points each) as a training data set and 600 points (c_1, c_1', c_2, and c_3: 150 points each) as a test data set and then merged c_1 and c_1' into a single class c_1 in each data set (Fig. 2a shows the test data set). Because this data set produces an apparently inseparable target, {1} versus {2,3}, by a simple classifier, poor classification performance would be expected when B^{1R} is used as the ECOC code matrix. When employing B^{AA}, which includes B^{1R} as the initial coding, the weight (confidence) optimized by our WMAP for such an unreliable target as {1} versus {2, 3} should shrink to a small value.

We constructed a total of six combinations of two multiclass classification algorithms, MAP and WMAP, and code matrices, 1R, 11, and AA. As an individual binary classifier, we used an SVM with a linear kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}'$. We set $\gamma_0 = 2$ and $\beta = 2,000$ for the (hyper)parameters.

These combinations were evaluated by the means and standard deviations of the three-class classification

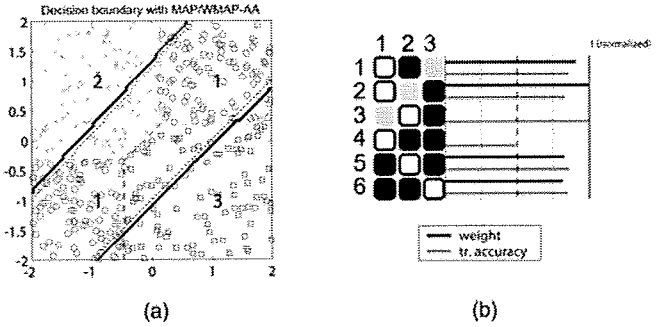

Fig. 2. Application of WMAP-AA and MAP-AA to a synthesized data set. (a) The scatter plot of the data points and the decision boundaries estimated by WMAP-AA (the solid lines) and by MAP-AA (the dotted lines). The dash lines represent the Bayes optimal decision boundaries. (b) The weight optimization result by WMAP-AA. The matrix represents AA coding of the three class problem (the symbols are same as those in Fig. 1). Two bars located in the right of each target vector represent the weight of the target (black: value normalized so that its maximum value became 1) and the training accuracy of the binary classifier for the target (gray).

accuracies for the training and test data sets over five different and random data generations (Table 1). As expected, WMAP-AA represented the best test accuracy among all combinations. The effect of the weight optimization of binary classifiers is remarkable; WMAP-11 and WMAP-AA showed higher test accuracies than MAP-11 and MAP-AA, which employ uniform weights. The reason why WMAP-11 had slightly lower test accuracy than MAP-11 could be overtraining. The performance of WMAP-1R was significantly improved compared to MAP-1R; the poor performance of MAP-1R was caused by a nuisance target in B^{1R}, {1} versus {2, 3}, which could not be discriminated well by a linear kernel SVM, and its weight successfully became almost zero in WMAP-1R.

The WMAP-AA result above can be seen as an example of how weight optimization by WMAP worked. To construct an optimal decision boundary by the whole multiclass classifier, it is better to ignore unreliable binary classifiers and also to appropriately weigh binary classifiers so as to contribute to the final multiclass classification performance; seeking effectively an appropriate "graded" coding starting from the original coding, in this case, B^{AA}. This experiment demonstrated that our WMAP-AA automatically meets this requirement. Fig. 2 shows the decision boundary (Fig. 2a) and the weights of the binary classifiers (Fig. 2b) obtained by WMAP. Interestingly, the weights of targets {1} versus {2, 3} and {2} versus {3} became 0 in this result. Since it is difficult to train the binary classifier for the target {1} versus {2, 3} (Fig. 2b, line 4), thus making it

TABLE 1
Classification Performance of Combinations of Binary Classifiers for an Artificial Problem

	MAP-1R	MAP-11	MAP-AA
Training	0.5625 (0.0643)	0.8785 (0.0088)	0.8415 (0.0297)
Test	0.5567 (0.0497)	0.8687 (0.0155)	0.8313 (0.0152)
	WMAP-1R	WMAP-11	WMAP-AA
Training	0.8670 (0.0330)	0.8825 (0.0105)	0.8925 (0.0173)
Test	0.8603 (0.0211)	0.8637 (0.0197)	**0.8783** (0.0130)

TABLE 2
Gene Expression Data Sets of Four Tumor Classification Problems

Dataset	# of samples	# of classes	# of genes
Thyroid cancer	168	4	2,000
Esophageal cancer	141	3	1,763
SRBCT	83	4	2,308
Leukemia	72	3	11,225

unreliable, the weight of {1} versus {2, 3} became approximately zero to ignore this classifier in the whole multiclass classifier. On the contrary, the weight of {2} versus {3} became approximately zero for another reason. The data points of classes c_2 and c_3 for target {2} versus {3} (Fig. 2b, line 3) have easily separable distributions, and we obtained a good binary classification performance for this target. If we put trust in this target, {2} versus {3}, however, the performance of the multiclass classification may degrade because c_1 would be classified into c_2 or c_3 randomly, based on the decision boundary for this target. The entire multiclass classifier preferred to emphasize other binary classifiers to achieve higher accuracy for c_1. The decision for c_2 and c_3 was then compensated by voting by other classifiers such as {2} versus {1, 3} and {3} versus {1, 2}. As a consequence of grading each element in the code matrix B^{AA} by optimizing the weights, the decision boundary by WMAP-AA came to have an expanded margin to minimize classification loss, in comparison to MAP-AA. The result of this simple artificial problem suggests that the optimal coding problem in ECOC can be solved by our weight optimization method (WMAP) by making unnecessary targets in the initial code matrix shrink.

5.2 Experiment 2: Applications to Tumor Classification Problems

Our method was next applied to four tumor classification problems based on gene expression profiling. The information of the data sets is summarized in Table 2, and the details are described below.

5.2.1 Thyroid Cancer Data Set

The thyroid cancer data set is composed of original gene expression profiles from four tissue types of human thyroid origin that contain 168 samples and 2,000 genes measured by an adaptor-tagged competitive PCR (ATAC-PCR) [18] method. The main diagnostic procedure for thyroid cancer is fine needle aspiration, but because the tissue structure is disrupted during the sampling process, differential diagnosis is extremely difficult [19], [20], [21]. Thus, diagnosis from gene expression profiles has been anticipated, though it would not be an easy task. The composition of the samples are 58 (follicular adenoma: FA), 28 (follicular carcinoma: FC), 40 (normal: N), and 42 (papillary adenocarcinoma: PC).

5.2.2 Esophageal Cancer Data Set

This data set is also composed of original gene expression profiles obtained from esophageal cancers of Japanese patients by ATAC-PCR [22], [23]. It should be noted that esophageal cancers in Japan are mostly squamous cell carcinoma, while those in the US and Europe are adenocarcinoma, i.e., Barret tumors. The task here is differential

diagnosis of three histological types: poorly differentiated (the sample number is 14), moderately differentiated (97), and well differentiated (30).

5.2.3 SRBCT Data Set [2]

Gene expression profiles about small round blue cell tumors (SRBCTs) of childhood, which contain 83 samples and 2,308 genes measured by cDNA microarrays, can be accessed at http://research.nhgri.nih.gov/microarray/Supplement/. SRBCTs, which include the Ewing family of tumors (EWS), rhabdomyosarcoma (RMS), Burkitt lymphoma (BL), and neuroblastoma (NB), have some difficulty in being distinguished solely histologically due to their similar appearance. The composition of the samples is 29 (EWS), 25 (RMS), 11 (BL), and 18 (NB).

5.2.4 Leukemia Data Set [24]

Gene expression profiles about three types of leukemia, which contain 72 samples and 11, 225 genes measured by Affymetrix oligonucleotide arrays, can be accessed at http://www-genome.wi.mit.edu/cancer. The composition of the samples is 28 (acute myeloid leukemia: AML), 24 (acute lymphoblastic leukemia: ALL), and 29 (*MLL* translocation: MLL).

We prepared the six ways of aggregating binary classifiers identically to those used in Experiment 1. For each binary classifier to be aggregated, we prepared an SVM with a linear kernel using all genes without any selection procedure. It should be noted that in many classification problems based on gene expression profiling, employing linear kernels in SVMs has exhibited better performance than employing more complicated kernels; since complicated kernels implicitly assume high-dimensional feature spaces, they may overfit the relatively large noise involved in gene expression data. We preset the (hyper)parameters of the MAP and WMAP methods at $\gamma = 2$ and $\beta = 2,000$ for the thyroid cancer, esophageal cancer, and SRBCT data sets, and at $\gamma = 2$ and $\beta = 1,500$ for the leukemia data set. We also prepared three state-of-the-art multiclass classification algorithms: a nearest shrunken centroid algorithm[2] (NSC) [25] and two direct implementations of MC-SVM, Weston and Watkins (WW) [13], and Crammer and Singer (CS), which is a modification of the WW approach [14]. These methods cast multiclass categorization problems as a constrained optimization problem with a quadratic objective function by introducing a generalized notion of the margin into multiclass problems.

In NSC, the shrinkage parameter Δ was optimized by searching from 0 to 6 at intervals of 0.25. In the two MC-SVM variants, a linear kernel was also employed, because it showed the best performance. The parameters for NSC and MC-SVM were optimized based on just the training data sets for avoiding information leak from the test data sets.

For each data set and each method, training accuracies and test accuracies were evaluated with a fivefold cross-validation framework, where for each split of the five folds, the ratios of all classes were maintained to be similar to the other folds. The mean and standard deviation of the results are shown in Table 3. For the SRBCT data set, all classifiers exhibited 100 percent accuracy at both training and test; so, the results are not shown in the table.

2. NSC is known as a method implemented in the PAM: Prediction Analysis for Microarrays software package (http://www-stat.stanford.edu/~tibs/PAM/).

TABLE 3
Cross-Validation Accuracies for Three Real Gene Expression Data Sets

	MAP-1R	MAP-11	MAP-AA
Thyroid Cancer			
Training	1 (0)	1 (0)	1 (0)
Test	0.762 (0.065)	0.762 (0.072)	**0.774** (0.074)
Esophageal cancer			
Training	0.821 (0.109)	0.901 (0.003)	0.901 (0.003)
Test	0.695 (0.026)	0.688 (0.076)	0.696 (0.050)
Leukemia			
Training	1 (0)	1 (0)	1 (0)
Test	**0.985** (0.033)	0.956 (0.068)	0.969 (0.069)
	WMAP-1R	WMAP-11	WMAP-AA
Thyroid Cancer			
Training	1 (0)	1 (0)	1 (0)
Test	0.762 (0.065)	0.762 (0.072)	**0.774** (0.074)
Esophageal cancer			
Training	0.901 (0.003)	0.917 (0.037)	0.901 (0.003)
Test	0.695 (0.026)	0.688 (0.076)	**0.703** (0.051)
Leukemia			
Training	1 (0)	1 (0)	1 (0)
Test	**0.985** (0.069)	0.956 (0.068)	0.969 (0.069)
	NSC	MC-SVM (WW)	MC-SVM (CS)
Thyroid cancer	($\Delta = 0.50$)		
Training	0.887 (0.022)	1 (0)	1 (0)
Test	0.744 (0.035)	0.768 (0.069)	0.762 (0.068)
Esophageal cancer	($\Delta = 0.00$)		
Training	0.912 (0.025)	1 (0)	1 (0)
Test	0.675 (0.051)	0.681 (0.056)	0.673 (0.065)
Leukemia	($\Delta = 0.00$)		
Training	0.972 (0.020)	1 (0)	1 (0)
Test	0.890 (0.057)	**0.985** (0.034)	0.969 (0.069)

Comparing the proposed six ways, the B^{AA} coding was often found to be better than the others; it was the best for the thyroid and esophageal data sets and comparable to the best for the SRBCT and leukemia data sets. For the three data sets except esophageal, the training CV accuracy by all of the six combinations reached the upper limit of 1.0. The SRBCT data may be too easy to be classified perfectly even for the test, while for the thyroid and leukemia data sets, the training CV accuracies of 1.0 might come from overfitting because the test CV accuracies did not reach 1.0.

Either of the two cases above can be a hazard to our WMAP procedure, because the training accuracy is so saturated that the room for the weight optimization is restricted. This is why the test CV accuracies were the same between WMAP and MAP in some cases. Even in such saturated cases, however, the optimization with respect to the soft-max accuracy can improve the aggregation of multiple binary classifiers, especially when there are a lot of constituent binary classifiers, as can be seen in Section 6. Compared to the existing state-of-the-art multiclass classification methods, we found our proposed methods, especially with weight optimization (WMAP), exhibited better or comparable performance.

5.3 Experiment 3: Applications to a Larger Class Problem

When the number of classes (M) is large, the initial setting of the exhaustive coding (AA) becomes computationally