Bull Yamaguchi Med School 53(3-4):37-43, 2006

The Development of a Novel Method for the Classification of the aCGH Profiles Based on Genomic Alterations

Motonao Nakao, Takae Okada, Kenzo Ikemoto, Tomoko Furuya, Atsunori Oga, Shigeto Kawauchi and Kohsuke Sasaki

Department of Pathology, Yamaguchi University Graduate School of Medicine, 1-1-1, Minami-Kogushi, Ube, Yamaguchi 755-8505, Japan
(Received November 20, 2006, accepted December 19, 2006)

Abstract Due to recent advances in array technologies, array-based comparative genomic hybridization (aCGH) is now widely used to detect DNA copy number aberrations in cells in the biomedical research field, especially in the field of oncology. However, it may be difficult to make reproducible and consistent categorizations of aCGH profiles. The development of a convenient and reliable classification method for the aCGH profiles is therefore necessary for the clinical application of the aCGH technology. We recently developed a novel classification method for the aCGH profiles based on genomic alterations in a reproducible fashion. This method has two steps, namely adaptive-weights-smoothing (AWS) and then a self-organizing map (SOM), which thus allows for the automatic classification of the aCGH profiles based on similarities in the genomic aberration pattern. We applied this method to 32 colorectal adenocarcinomas to demonstrate its practical utility. This method may thus be useful for identifying the genomic aberrations implicated in cancer phenotypes and clinicopthological correlations, because this method can be used for all aCGH data.

Key words: array CGH, colorectal cancer, classification, AWS (Adaptive Weights Smoothing), SOM (Self Organizing Map)

Introduction

Cancer classification based on genetic alterations is useful for the precise diagnosis and effective treatment of individual patients. Array technologies are practical for the comprehensive detection of genetic aberrations, and array-based comparative genomic hybridization (aCGH)[1)2)] is one of the representative array technologies which is also effective for detecting DNA copy number aberrations in the whole genome. A statistical analysis of the aCGH data is further needed to obtain clinically useful information. Since the size of the aCGH is large in comparison to the number of cases examined by aCGH, it is therefore difficult to apply conventional methods to the aCGH data sets formatted by different methods. Diverse statistical ways have been used to overcome such difficulties and to classify the relationship of genomic aberrations to clinicopathological information, even though such methods are technically complicated and are not always applied to all data sets. Therefore, another analytical method, which can be applied to all aCGH data sets formatted by different methods, is thus required to make it possible to classify the aCGH profile patterns based on genomic alterations. In the present study, we developed a novel method to meet these requirements for the classification of aCGH profiles. This method consists of two components, namely AWS (Adaptive-Weights-Smoothing) and SOM(Self-Organizing-Map). This method allowed for the classification of aCGH profiles based on the DNA copy number aberrations in tumor cells. In this paper, we exem-

plified the practical utility of this method by applying the method to aCGH data obtained from 32 colorectal adenocarcinomas.

Materials and methods

1. aCGH

aCGH was performed using the MacArray Karyo4000 (Macrogen Inc, Seoul Korea). The array consisted of duplicate spotted 4030 human bacterial artificial chromosome (BAC) clones that covered the human whole genome at an average interval of 1Mbp. We used 32 surgically removed sporadic colorectal adenocarcinoma specimens for aCGH. The patients consisted of 15 males and 17 females (Table 1). High molecular weight genomic DNA was extracted from microdissected tumor tissue spe cimens using a DNA extraction kit. (Dneasy Tissue kit, QIAGEN SCIENCES) [3)4)] Sample DNA (500ug, BioPrime DNA Labeling System Invitrogen) was labeled with Cy5. Reference DNA (Human Genomic DNA Promega) is labeled with Cy3. These DNAs were applied to a CGH array. Images of 16bit fluorescence intensity were captured from each array with a Gene Pix 4000A scanner (Axon Instruments, Burlingame, CA). Any spots with an obviously weak signal intensity were usually excluded from the analysis as a missing value. Such a spot was seen at a rate of a few percent, which varied depending on the quality of the array, the performance of the scanner, and the degree of experimental error. The ratio of the fluorescence intensity for each clone was calculated and then converted to log base 2. The values were normalized using LOWESS (Locally Weighted Scatter plot Smoother) because the dynamic range of the two fluorescence dyes is different. If the experiment progresses ideally, then the log2 ratios for each case would be median centered to Zero (log2) in a case without any DNA copy number aberrations.

Table 1 Statistics analysis condition

SOM analysis condition
Topology : rectangular
Rows x columns : 5 x 5
Type of the kernel function : Gaussian
Initial radius / final radius : 2.0 / 0.1
Count of cycles : 100
k-means method condition
Cluster : 25
Count of cycles : 10,000
Error threshold : 10E-5
Sample condition
number : 32 cases
C001 – C034 (C018, 24 are absent)
15 male, 17 female

2. AWS

Adaptive Weights Smoothing (AWS) was applied to the ratio data obtained. We used the GLAD program developed by R package.[5)6)] R is a language and environment for statistical computing and graphics. The AWS processing process of the original data is shown in Fig. 1. Three consecutive points are needed to estimate each data point with AWS. The value of each point may become invalid, but smoothing is completed, and thereafter we can see the CGH profile clearly.

3. SOM (*Self-Organizing Map*)

Furthermore, the CGH profiles processed with AWS were classified by means of SOM. The SOM analysis was performed using a program proposed by R package. The SOM analysis was performed three times with a rectangle grid measuring 5 row 5 columne. In addition, the calculation cycles were fixed at 100 times in this experiment. The SOM analysis was calculated by the Eulidean distance.

4. k-means analysis

The k-means analysis method was applied to the CGH profiles treated by AWS as well as the SOM method. As in the case of the SOM analysis, the cluster is at 25 and the

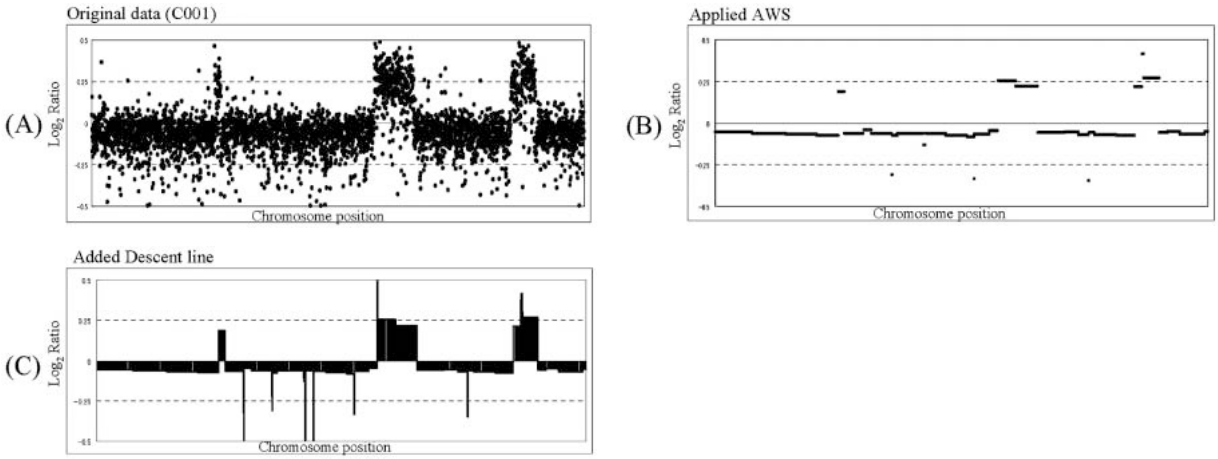

Fig. 1 (A) A representative profile (Case C001) of original data. Fluorescence intensity ratios (log base 2) of tumor DNA versus the reference DNA are shown for each BAC clone. DNA copy number aberrations are detected as clusters of BAC clones at several chromosomes such as 4, 11, 12, 19, and 20, though single BAC clone aberration scattered in the entire genome makes it difficult to identify break points. (B) An application of AWS (Adaptive Weights Smoothing) to the original data converts the complex profile of original data set to a simple profile pattern in which DNA copy number is described as a line. The interrupted parts of the line indicate break points in this profile. As a result, the chromosomal regions with DNA copy number aberrations are exaggerated by the AWS treatment of the original data set. (C) Descent lines to Zero point are added to the line in the profile with AWS treatment to clarify the degree of the DNA copy number aberrations. It is possible to measure the degree of DNA copy number aberrations for each chromosome. Abscissa: chromosomal number (Form the 1st left to 22,X and Y). Ordinate: DNA copy number ratio (log2).

calculation cycles are 10,000 times. The k-means analysis was calculated by the Eulidean distance.

Results and discussion

The aCGH profile patterns representing the DNA copy number gains and losses were compared before and after treatment with AWS in 32 colorectal adenocarcinomas. It was evident that the chromosomal regions with the DNA copy number aberrations became clear after the AWS treatment (Fig. 2). This was why the smoothing of the aCGH data profile by the AWS exaggerated the DNA copy number gains and losses in the profile pattern. The background fluctuations were depicted as a flat line in the aCGH profile, while weak aberrations in the DNA copy number were depicted as a concavity or convexity line. As a result, the chromosomal regions with DNA copy number aberrations were easily identified in the aCGH profile.

After treatment to determine the aCGH profile, either the k-means method or SOM was applied to the data and the calculations were repeated three times. Although 32 profiles were classified into 25 node by each method (Fig. 3), the convoluted patterns were different between these analytical methods. The aCGH profiles were classified by the k-means or SOM methods. With the k-means method, 16 to 18 aCGH profiles were classified into a single node, and eventually 32 aCGH profiles were classified into seven or eight groups. Each group contained one, two or three profiles. With the SOM analysis, on the other hand, five profiles were not classified into a single node, and 32 profiles were classified into 17 to 19 groups. The present results indicate that the classification performance was higher for the SOM method

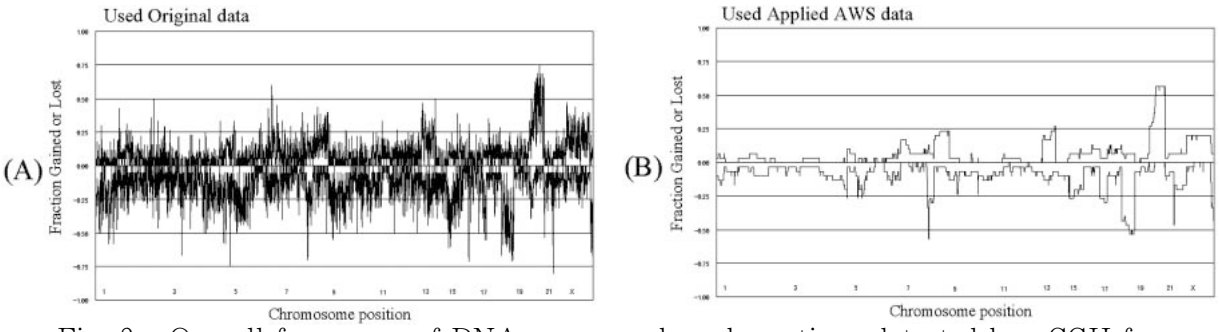

Fig. 2 Overall frequency of DNA copy number aberrations detected by aCGH for each BAC clone in 32 colorectal adenocarcinomas. The frequency of aberrations is depicted as a fraction of cases with DNA copy number gain or loss for 4000 BAC clone (the entire genome). The dots in the upper part of the profile indicate the frequency of tumors with DNA copy number gains, and dots in the lower part of the profile indicate the frequency of tumors with DNA copy number losses. (A) A histogram is formed by using the original data set. The frequency greatly varies from clone to clone. It is difficult to differentiate between the chromosomal regions with highly frequent aberrations and those with infrequent aberrations. (B) A histogram is formed by using data sets treated with AWS. The histogram makes it easy to identify the chromosomal regions with and without clusters of DNA copy number aberrations.

(A)

2	0	3	0	2
1	0	1	1	1
0	3	0	1	1
1	0	1	1	1
4	1	1	2	4

2	2	1	0	3
2	0	0	2	1
1	0	1	1	2
1	1	0	0	1
3	1	2	2	3

4	0	1	2	2
0	3	2	1	1
2	2	0	0	2
0	0	2	2	0
1	1	0	1	3

(B)

0	0	0	0	18
0	0	0	0	0
0	0	0	0	0
0	0	0	0	1
1	3	1	5	3

0	0	0	0	16
0	0	0	0	0
0	0	0	0	0
0	0	0	1	1
3	1	5	3	2

0	0	0	0	0
0	0	0	19	0
0	0	0	0	1
0	0	1	0	0
2	0	1	1	7

Fig. 3 Grids of SOM (A) and k-means (B). Both SOM and k-means methods classify the profiles of 32 tumors into 5 x 5 cells. The figures in the cells of the grid indicate the number of profiles that are included in the same node. SOM and k-means methods are three times applied to the data obtained by the AWS treatment of the original data, producing three grids for SOM and k-means methods. These three grids for each analysis method are similar to each other. In each analysis, tumors with a similar profile pattern are included in the same cell. In the SOM analysis, the degree of similarity of profile patterns between tumors depends on the distance between the cells. For the k-means the distance between the cells and the similarity of aCGH profile are unrelated.

than for the k-means method. Furthermore, SOM analyses allowed for the easy identification of tumors with similar genomic profiles.

It is true that with the increase in the number of cases examined, it becomes difficult to assess the difference and similarity of the aCGH profile patterns between tumors. A large amount of data (variables are more

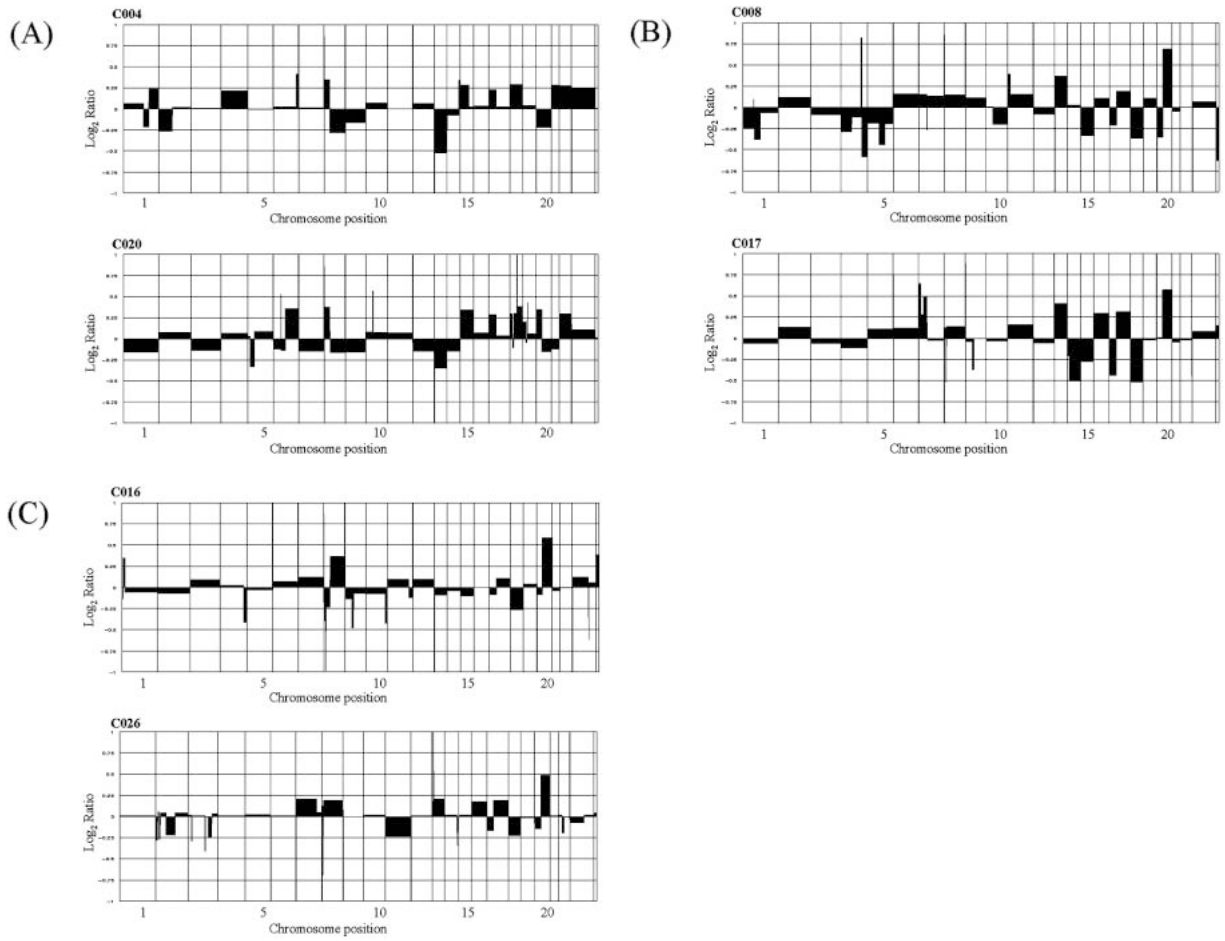

Fig. 4 Examples classified by the SOM analysis. A pair of tumors show a similar profile pattern. These two tumors are classified into the same cell by the SOM analysis 3 consecutive times. Each pair of profiles in A, B, and C are judged to be similar each other by the method developed in this study. It is difficult to classify the CGH profiles of cancer based on the appearance, the present analytical method allows for the classification of CGH profiles according to the similarity of profile patterns. (A): In tumors of C004 and C020, both CGH profiles show 8p gain, 8q loss, 13 loss, 15 gain, and 22 gain. (B): In tumors of C008 and C017, the CGH profiles show 13 gain and 15 loss, 17p loss, 17q gain, 18 loss and 20q gain. (C): In tumors of C016 and C026, the CGH profiles a show18 loss and a 20q gain. Abscissa: chromosomal position (Form the 1st left to 22,X and Y). Ordinate: DNA copy number ratio (log2).

than thousands) are provided by array technologies such as genome array, expression array,[7] SNP array,[8] and protein chips.[9] In this situation, we inevitably convert the experimental data into information linked with the clinicopathologic data when using the array data for both the diagnosis and treatment of each cancer patient. In this context, SOM is superior to the k-means method for distinguishing tumors with a similar genome aberration pattern from others with different patterns in a large-scale study, because the k-means method and hierarchical clustering analyses can hardly differentiate tumors with a similar profile from others. The sophisticated method presented herein is appropriate for the aCGH data analysis from a clinical point of view. To our knowledge, this is the first attempt to classify the CGH profile pattern by the method combining AWS and SOM. Fig. 4(A), 4(B), and 4(C) show the nodes classified by SOM. Each profile in Fig.

4 was classified into the same node based on similarities of the profile pattern, though the magnitude of the DNA copy number aberration in these profiles was apparently different. Even though the tumors C016 and C026 (Fig. 4(C)) were classified into the same node, the DNA copy number aberrations in C016 still seemed to be much more than those observed in C026.

The present data-mining method is thus considered to be useful for tumor genetic analysis regarding not only colorectal carcinogenesis but also tumor characterization. In general, cancers have been classified into a few groups for the precise diagnosis and patient management, and they are also categorized based on genomic alterations including. The total number of the aberrations,[10] the frequency in aberrations of some specific chromosomal regions[11] or the magnitude of genome instability.[12] Such genomic alteration data have the potential to predict the patient's prognosis,[13)14)] help in the discovery and development of drugs,[15] explore novel genes[16] or characterize genes.[17] In this series, we identified two tumors that showed a close resemblance of CGH profiles. These two tumors may develop through the same genetic pathway in colorectal carcinogenesis and they may also show similar biological characteristics. The categorization of the aCGH profiles of colorectal cancer may lead to the clarification of genomic pathways in the development and progression of sporadic colorectal cancers. This new method that makes it possible to categorize aCGH profiles may allow for the estimation of the patient prognosis and the selection of optimal treatment in individual patients. Further large-scale studies are thus needed to validate the clinical usefulness of this analytical method.

References

1) Pollack, J.R., Perou, C.M., Alizadeh, A. A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. and Brown, P.O.: Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.,* **23** : 41-46, 1999.

2) Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W. and Albertson, D.G.: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.,* **20** : 207-211, 1998.

3) Hashimoto, Y., Oga, A., Okami, K., Imate, Y., Yamashita, Y. and Sasaki, K.: Relationship between cytogenetic aberrations by CGH coupled with tissue microdissection and DNA ploidy by laser scanning cytometry in head and neck squamous cell carcinoma. *Cytometry,* **40** : 161-166, 2000.

4) Hashimoto, Y., Oga, A., Kawauchi, S., Furuya, T., Shimizu, N., Nakano, T., Imate, Y., Yamashita, H. and Sasaki, K.: Amplification of 3q26 approximately qter correlates with tumor progression in head and neck squamous cell carcinomas. *Cancer Genet. Cytogenet.,* **129** : 52-56, 2001.

5) Polzehl, J. and Spokoiny, V.G.: Adaptive Wei ghts Smoothing with applications to image restoration. *J. Roy. Stat. Soc. B,* **2** : 335-354, 2000.

6) Hupe, P., Stransky, N., Thiery, J.P., Radvanyi, F. and Barillot, E.: Analysis of array CGH data; from signal ratio to gain and loss of DNA regions. *Bioinformatics,* **20** : 3413-3422, 2004.

7) Schena, M., Shalon, D., Davis, R.W. and Brown, P.O.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science,* **270** : 467-470, 1995.

8) Gentalen, E. and Chee, M.: A novel method for determining linkage between DNA sequences; hybridization to paired probe arrays. *Nucleic Acids Res.,* **27** : 14 85-1491, 1999.

9) Hutchens, T.W. and Yip, T.T.: New desorption strategies for the mass spectro metric analysis of macromolecules. *Rapid Commun. Mass Spectrom.,* **7** : 576-580, 1993.

10) Blaveri, E., Brewer, J.L., Roydasgupta, R., Fridlyand, J., DeVries, S., Koppie, T., Pejavar, S., Mehta, K., Carroll, P., Simko, J.P. and Waldman, F.M.: Bladder cancer stage and outcome by array-based comparative genomic hybridization. *Clin.*

Cancer Res., **11** : 7012-7022, 2005.
11) Tanner, M.M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W.L., et al.: Increased copy number at 20q13 in breast cancer; defining the critical region and exclusion of candidate genes. *Cancer Res.,* **54** : 4257-4260, 1994.
12) Fridlyand, J., Snijders, A.M., Ylstra, B., Li, H., Olshen, A., Segraves, R., Dairkee, S., Tokuyasu, T., Ljung, B.M., Jain, A. N., McLennan, J., Ziegler, J., Chin, K., Devries, S., Feiler, H., Gray, J.W., Waldman, F., Pinkel, D. and Albertson, D.G.: Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer,* **6** : 96, 2006.
13) Suehiro, Y., Umayahara, K., Ogata, H., Numa, F., Yamashita, Y., Oga, A., Morioka, H., Ito, T., Kato, H. and Sasaki, K.: Genetic aberrations detected by comparative genomic hybridization predict outcome in patients with endometrioid carcinoma. *Genes Chromosomes Cancer,* **29** : 75-82, 2000.
14) Ueno, T., Tangoku, A., Yoshino, S., Abe, T., Toshimitsu, H., Furuya, T., Kawauchi, S., Oga, A., Oka, M. and Sasaki, K.: Gain of 5p15 detected by comparative genomic hybridization as an independent marker of poor prognosis in patients with esophageal squamous cell carcinoma. *Clin. Cancer Res.,* **8** : 526-533, 2002.
15) Plasencia, C., Rooney, P.H., Taron, M., Martinez-Balibrea, E., McLeod, H.L. and Abad, A.: Chromosomal imbalance maps of human 5FU-resistant colorectal cancer cell lines: implications in the analysis of 5FU-acquired resistance mechanisms. *Int. J. Oncol.,* **22** : 945-953, 2003.
16) Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringner, M., Sauter, G., Monni, O., Elkahloun, A., Kallioniemi, O.P. and Kallioniemi, A.: Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.,* **62** : 6240-6245, 2002.
17) Ghosh, M., Koike, N., Tsunoda, S., Hirano, T., Kaul, S., Kashiwagi, H., Kawamoto, T., Ohkohchi, N., Saijo, K., Ohno, T., Miwa, M. and Todoroki, T.: Characterization and genetic analysis in the newly established human bile duct cancer cell lines. *Int. J. Oncol.,* **26** : 449-456, 2005.