Bull Yamaguchi Med School 56(3-4):49-55, 2009

Identification of the Copy Number Aberrations for Determining the Disease Stage of Colorectal Cancers: The Application of a Multifactor Dimensionality Reduction (MDR) Method to an Array-Based CGH

Motonao Nakao and Kohsuke Sasaki

Department of Pathology, Yamaguchi University Graduate School of Medicine, 1-1-1
Minami-Kogushi, Ube, Yamaguchi 755-8505, Japan

(Received November 13, 2009, accepted January 8, 2010)

Abstract The clinical application of array CGH technology is eagerly awaited. It is necessary to clarify the genome markers closely associated with the clinical condition. The multifactor dimensionality reduction Method (henceforth, MDR) analysis was applied to the array CGH data of 74 cases colorectal cancer in order to identify two or more spot clones, i.e., a clone marker, with the maximum separation ability between low stage and a high stage lesions.

The optimal marker in one clone was 8q24.3 (Accuracy 0.7027, Sensitivity 0.9688, Specificity0.5000), The optimal markers in two clones were 7q36.3 and 22q11.1 (Acc. 0.8108, Sen. 0.8438, Spe. 0.7857). Moreover, the optimal markers in three clones are 7p22.2, 8p23.3, and 15q11.2 (Acc. 0.9054 Sen. 0.9375, Spe. 0.881). However, all the values of testing accuracy were very low, and it cannot be trusted as a clinical marker. When three clones (8p23.3, 8q21.11, 8q24.3) of only chromosome 8 were used as a marker, testing accuracy exceeded 85%. The large quantities of data from a microarray - the MDR method - can efficiently narrow down the number of clones. Moreover, since there was a large Copy Number Imbalance (CNI) in a chromosomal region unit in CGH, an analysis of a specific chromosome was efficient. An analysis of chromosome 8 was effective for the stage classifications of colorectal cancer.

Key words: array based CGH, multifactor dimensionality reduction (MDR) method, cross validation test

Introduction

The clinical application of the microarray technology is eagerly awaited, under current conditions, the data are not applicable. It is important to avoid the analysis of a huge amount of data. Therefore, it is difficult to apply a normal multivariate analysis to such data. The analysis is performed by various approaches.[1)2)] A normal multiple comparison determines whether there is a difference in every case with regard to a specific variable or variable quantities, and the significance level of each variable is calculated. Therefore, a microarray can give a lot of false-positive results. Hypotheses on a set of data using the Bonferroni correction or the Holm-Bonferroni method post-test are applied,[3)] but a microarray frequently yields too many variables, so that the threshold of the P-value that applies the correction is too severe and lowers the power of the test greatly. As a result, the true-negative level increases. Because the correction assumes that each variable is independent, it cannot be applied in single nucleotide polymorphisms (SNPs) and copy number imbalance (CNI) that is a mutation absorbed on a series of chromosome.

One purpose of the CGH clinical application is to identify biomarkers that predict a patient's prognosis and drug resistance, etc. based on specific chromosome clones. A marker of liver cancer was identified that was suiTable for differentiating a low stage (stage I / II) from a high stage (stage III / IV) among each of the 1440 clones[4)] as well as a marker of stomach cancer[5)] and another that could identify the existence of lymph node metastasis of colorectal cancer.[6)] The one clone with the highest separation of low and high stage of colorectal cancer was 8q24.3 (Accuracy 0.7027, Sensitivity 0.9688, Specificity 0.5000). However, a more reliable classification could be attained by combining two or more markers. The use and improvement of the MDR (multifactor dimensionality reduction) method has been advanced for SNPs with a large amount of data. MDR is a nonparametric and genetic model-free alternative to logistic regression for detecting and characterizing nonlinear interactions among discrete genetic and environmental attributes. The MDR method combines attribute selection, attribute construction, classification, cross-validation, and visualization to provide a comprehensive and powerful data mining approach to detecting, characterizing, and interpreting nonlinear interactions. Originally the MDR method analysis has been utilized in a SNPs study. The current study investigated the application of this method to array CGH, and also differentiated each case into a low stage and a high stage, and therefore the combination of the best clones was thus examined in the present study.

Materials and methods

A total of 74 colon cancers were examined (Table 1). These cases were divided into two groups for convenience: stages I/II (31 cases) and stage III/IV (42 cases). An array-based CGH was conducted for these cases using a previously reported procedure. Specifically, DNA (500 ng) extracted from a cancer cell that had been selectively isolated from a cancer tissue with tissue microdissection was labeled with Cy 3 (Perkin Elmer, Wellesley, MA) and control DNA with Cy 5 (Perkin Elmer). Using an array (Macrogen, Korea), in which 4030 BAC clones were spotted as duplicates in the presence of Cot-1 (50 mg, Gibco BRL, Gaithersburg, MD), hybridization was conducted for 72 hr at 37oC. A GenePix 4000A scanner (Axon Instruments, Union City, CA) was used for reading the fluorescence signals. The ratio of the fluorescence intensity of Cy3/Cy5 was recorded with a two-bottom log. The Log2 fluorescence intensity ratio was the raw data computed from a CGH set 0.25 or more to "gain" and the portion where sets less than -0.25 to "loss" and the state where it is normal in between, and a statement sets to "0" and which does not have data made the statement "NA". MDR is available as an open-source (GPL) software package. It is a cross-platform program written entirely in Java. It is available from the MDR web site [http://www.multifactordimensionalityreduction.org/]. The workstation which executed this program is Windows-XP professional sp2, Jave5.0SE1.5

Table 1 Clinical data of 74 colorectal cancer

Sex	male	40
	female	34
location	Cecum, Ascending colon Transverse colon	22
	Descending colon, Sigmoid colon	21
	Rectosigmoid, Rectum, Proctos	31
Stage	I or II	32
	III	28
	IV	14

The J48 and LMT classification by the WEKA program were performed for all 4030 clones as comparative experiments, with raw FIR data(Fig. 1, Fig. 2). The MDR method analysis for all the 4030 clones was conducted subsequently. Since the MDR method required computation time, using all 4030 clones, it has calculated only up to two clones. Therein, the MDR method analysis was conducted using the clone used as P< 0.05 in chi-square test (67 clones). Thereafter, the MDR method was analyzed by using the clone whose frequency of CNI in 74 all cases

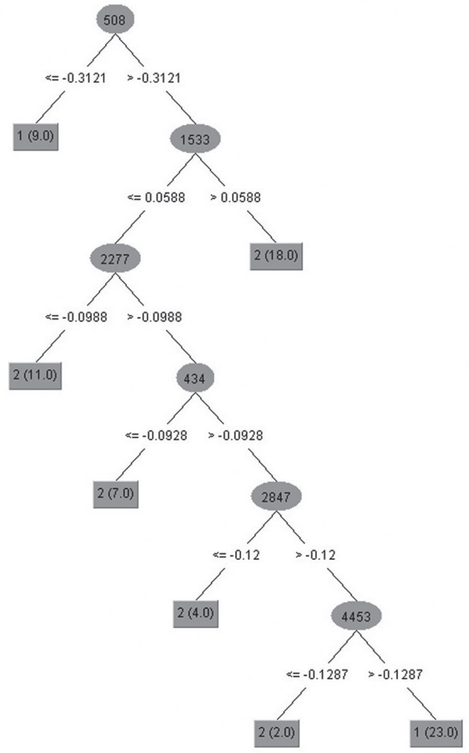
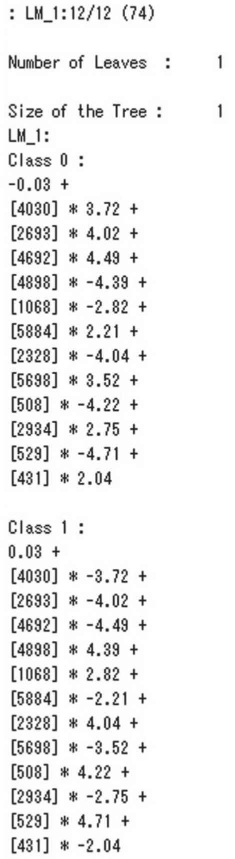


Fig. 1 J48 division tree model result. The symbol 'terminal' of a flow chart figure shows a clone name, and 'processing' shows the class name classified. Class 1 is low stage(stage I&II) and class 2 is high stage(stage III&IV). Clone 508 is located 19p13.2. Clone 533, 2277, 434, 2847, 4453 is located each Xp22.2, 12q24.11, 9p24.1, 12p13.31, 3p24.3.



```
: LM_1:12/12 (74)

Number of Leaves  :     1

Size of the Tree :      1
LM_1:
Class 0 :
-0.03 +
[4030] * 3.72 +
[2693] * 4.02 +
[4692] * 4.49 +
[4898] * -4.39 +
[1068] * -2.82 +
[5884] * 2.21 +
[2328] * -4.04 +
[5698] * 3.52 +
[508] * -4.22 +
[2934] * 2.75 +
[529] * -4.71 +
[431] * 2.04

Class 1 :
0.03 +
[4030] * -3.72 +
[2693] * -4.02 +
[4692] * -4.49 +
[4898] * 4.39 +
[1068] * 2.82 +
[5884] * -2.21 +
[2328] * 4.04 +
[5698] * -3.52 +
[508] * 4.22 +
[2934] * -2.75 +
[529] * 4.71 +
[431] * -2.04
```

Fig. 2 LMT division tree model result. LMT builds the tree of a logistic model. The used clone is as follows. Clone 4030 is located 1p31.1. Clone 2693, 4692, 4898, 1068, 5884, 2328, 5698, 508, 2934, 529, 431 is located each 1q21.3, 2q33.1, 3p14.1, 5p15.32, 9p24.3, 10q24.32, 12q24.33, 14q12, 17p12, 19q13.33, 20p12.1.

is 30% or more (819 clones). Finally, for time shortening of the MDR method, we used only chromosome 8 frequency of CNI used 30% or more (172 clones) of the clones. Chromosome 8 is high appearance frequency of the previous result of MDR method.

Results

Using WEKA classifiers trees J48 and LMT

A training set was conducted with the classification machine by WEKA and the results are shown in Fig. 1 (J48), and Fig. 2 (LMT). The classification accuracy of J48 is 100% (inside of 74 examples), and the classification accuracy of LMT is 97.3% (72 examples are Correctly Classified Instances among 74 examples). However, in the result of Cross-validation by 20 folding, J48 became 37.84% (inside of 74 examples 28 examples) LMT became 40.5% (inside of 74 examples 30 examples). The numerical value implies that the reliability of the result is remarkable low.

Using All 4030 clones

Clone 2748 (8q24.3) was the optimal solution in one clone distinguishes advanced cancer and early cancer (Accuracy 0.7027, Sensitivity 0.9688, Specificity 0.5000) (Table 1). However, the MDR method analysis of 4030

Table 2 MDR result of all 4030 clone

clone No.	Training Accuracy	Training Sensitivity	Training Specificity	Testing Accuracy	Cross-validation Consistency
2748	0.7027	0.9688	0.5000	0.6712	9/10
2130, 5109	0.8108	0.8438	0.7857	0.5479	2/10

clone No.	Cyto	Bac start	Bac end	Gene
2748	8q24.3	146119326	146201248	ZNF16, TMED10P, C8orf77,
2130	7q36.3	158569297	158654736	VIPR2, LOC644525, LOC729057,
5109	22q11.1	14458245	14564930	DUXAP8, LOC400879, LOC441969,

Table 3 MDR result of 67 clone (Chi-squar test p<0.05)

clone No.	Training Accuracy	Training Sensitivity	Training Specificity	Testing Accuracy	Cross-validation Consistency
2748	0.7027	0.9688	0.500	0.6712	9/10
2748, 2704	0.7973	0.9688	0.6667	0.6667	4/10
5579, 2031, 4494	0.8649	0.9688	0.7857	0.6324	2/10
2748, 5579, 5646, 2389	0.9189	1.000	0.8571	0.7627	6/10
2748, 5579, 5917, 2389, 4494	0.9595	1.000	0.9286	0.5918	2/10
2748, 5579, 4214, 4289, 5785, 2389	0.9595	1.000	0.9286	0.4750	2/10

clone No. The attributes that participated in the best model discovered.
Cross-validation 10th of intervals into whitch to divide the data, for the purpose of cross-validation.

clone No.	Cyto	Bac start	Bac end	Gene
2748	8q24.3	146119326	146201248	ZNF16, TMED10P, C8orf77,
2704	11p15.4	8182650	8272559	LOC644497, LMO1,
5579	8p23.3	649638	867290	ERICH1, C8orf68, LOC401442,
2031	14q32.12	91466488	91625006	FBLN5, TRIP11, PTMAP7, ATXN3,
4494	10q21.3	65365356	65466873	
5646	12q24.33	131529697	131689973	KIAA1545, LOC645277,
2389	14q32.33	104557112	104662718	CDCA4, GPR132,
5917	17p13.3	2510283	2624227	PAFAH1B1, KIAA0664,
5785	20p13	311329	442419	TRIB3, RBCK1, TBC1D20, CSNK2A1,
4214	9p21.3	21726099	21856741	LOC402359, MTAP,
4289	9q33.1	119261756	119362026	

Table 4 MDR result of 819 clone (Frequency >0.3)

clone No.	Training Accuracy	Training Sensitivity	Training Specificity	Testing Accuracy	Cross-validation Consistency
2748	0.7027	0.9688	0.5000	0.6986	20/20
2130, 5109	0.8108	0.8438	0.7857	0.5714	10/20
912, 5579, 5258	0.9054	0.9375	0.8810	0.5333	8/20

Cross-validation: 20th of intervals into whitch to divide the data, for the purpose of cross-validation.

clone No.	Cyto	Bac start	Bac end	Gene
912	7p22.2	2809992	2933327	GNA12, CARD11,
5579	8p23.3	649638	867290	ERICH1, C8orf68, LOC401442,
5258	15q11.2	18881005	18958308	LOC283755,

clones showed that optimal solutions in the combination of two clones are 2130 and 5109 (Acc. 0.8108, Sen. 0.8438, Spe. 0.7857). The independent separation ability of clone 2130 is the low stages 10/31 and the high stages 17/40, and of clone 5109 is the low stages 12/31 and the high stage 22/40. There is almost no difference between the two. When both clones were used, the separation ability increased to 81%. However, the testing accuracy fell to 0.5479. The upper number of clones still required more than two clones to perform the classifying of a low stage and a high stage at the actual clinical spot.

Using 67 clones passed chi-square test (p<0.05)

The result of the MDR method analysis to distinguish between high and low stages using 67 clones with a small p-value (<0.05) is shown in Table 3. When five clones and six clones were used from the analysis, the result in 67 clones showed a 96% classification ability, but the testing accuracy achieved a maximum of 0.7627 with four clones and the value fell as the number of clones increased to 5 and 6 clones after that. The optimal separation ability was combination of four clones in the analysis of 67 clones. The four clones suiTable for the stage determination were 2748, 5579, 5646, and 2389 (Acc. 0.919, Sen. 1.00, Spe 0.857).

Using 819 clones whose frequency of CNI was detected in more than 30% of cancers

Next, in order to identify the clone which uses three clones which make low and high stage classification the maximum, the MDR method was analyzed by proportioned about CNI using only the clone of 0.3 or more CNI frecuency in colorectal cancer (Table 4). In the case of one clone and two clones, it was the same result as shown in Table 2 which used all 4030 clones. The result used three selected clones (912, 5579, and 5258) is Acc. 0.9054, Sen. 0.9375, Spe. 0.881. In comparison of the optimal solution when using 67 clones, the degree of correctness improved drastically by using three clones, but testing accuracy is still 0.5333 as low. This fact means that the credibility of training accuracy of 3 clones is low as a result. However, in the present system, it is impossible of calculation of MDR method without reducing rather than 819 clone.

Using 172 clones on chromosome 8

Clones from chromosome No. 8 performed well in the MDR method analysis that maximizes the stage separation. The result of the MDR method analysis using only chromosome No. 8 clones is shown in Table 5. Although the testing accuracy was lowered at the time of 2 clones, while testing accuracy was set to 0.8548 at the time of three clones. It is important that testing accuracy rises statistically, the high value of testing accuracy means the high credibility, and we can expect the result by the examination in a real clinical field. Furthermore, the result of the cross validation test (20/20) was perfect. Three clones (5579, 4993, and 2748) shows Acc.0.8919, Sen.1.000, Spe.0.8095 with training test.

Table 5 MDR result of 172 clone (chromosome 8)

clone No.	Training Accuracy	Training Sensitivity	Training Specificity	Testing Accuracy	Cross-validation Consistency
2748	0.7027	0.9688	0.5000	0.6986	20/20
2748, 5579	0.7888	0.9671	0.6529	0.6667	13/20
5579, 4993, 2748	0.8919	1.0000	0.8095	0.8548	20/20
5579, 2978, 1310, 2748	0.9474	1.0000	0.9073	0.8545	17/20

Cross-validation: 20th of intervals into whitch to divide the data, for the purpose of cross-validation.

clone No.	Cyto	Bac start	Bac end	Gene
4993	8q21.11	77768566	77851358	ZFHX4,
2978	8q24.3	143892605	144004977	GML, LOC646338, CYP11B1, CYP11B2,
1301	8p11.21	42349308	42434978	LOC727725, DKK4, VDAC3, SLC20A2,

Discussion

Although a microarray can generate a lot of data, it is necessary to clarify genome markers closely connected with the clinical variable. The method for distinguishing a clinical variable using these data was established in previous studies. For example, a classification machine (J48, LMT) unfortunately yielded low testing accuracy and Cross-validation in microarray analyses with a few clinical sample numbers. The MDR method is used to raise these values to the maximum when determining a genome marker. The MDR method analysis for all the 4030 clones on the array could calculate only two clones within a limited time because there were too many clones. In addition, the testing accuracy in two clones was less than 60 percent. Therefore, it is necessary to significantly improve the testing accuracy in order to determine a general-purpose a genome marker. In addition, since the computational complexity increases exponentially for a variable , it is necessary for MDR method analysis to minimize the number of the clones in a large amount of data which were analyzed by microarray to find the most suiTable solution.

Then, the MDR was analyzed using only clones that strongly differentiated between a low stage and a high stage according to the chi-square test ($p<0.05$), as a general filter. Although this achieved a training accuracy > 90% in the combination of four clones, the maximum testing accuracy was 0.7627. Furthermore this value decreased if the number of clones was increased. If the variation of CNI is completely independent (random), this selection method is satisfactory. However, the actual clone is arranged on the chromosome and it is obviously influenced by the CNI of a clone on the same chromosome, especially when they are physically close.

Next, the 0.3 or more frequency of CNI was used to select items to conduct the MDR method analysis. The combination of 3 clones yielded a training accuracy >90%, but the testing accuracy was still remarkably low. In order to analyze an increasing number of clones, it is necessary to further extract the number of clones. Next, in order to limit the number of clones further, those on chromosome No. 8 were used. In this trial, only 3 clones of No. 8 chromosome yielded a training accuracy and testing accuracy of greater than 80%. In addition, the cross-validation test achieved the same result in 20 out of 20 trials, and if these clone are used, the classification of high stage and low stage in colorectal cancer can be expected to be 85% or more. This suggested that the loss of 8p23.3, gain of 8q24.3 and CNI of 8q21.11 were risk factors for high stage cancer. Moreover, a high stage prediction of colorectal cancer can be performed at a rate of 90 percent, using only the No. 8 chromosome.

Conclusion

The MDR method of Array CGH is a very effective to determine two or more genome markers associated with the clinical condition. However, since the computational complexity of the MDR method increased exponentially to the target variable, it is necessary to limit a specific chromosome.

References

1）Ashburner, M., Ball, C.A., Blake. J.A., et al.: Gene oncology: Toll for the Unification of biology. *Nat. Genet.*, **25**：25-29, 2000.

2）Harris, M.A., Clark, J., Ireland, A., et al.: The Gene oncology. *Nucleic Acids Res.*, **32**：D258-D261, 2004.

3）Holm, S.: A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**：65-70, 1979.

4）Chochi, Y., Kawauchi, S., Nakao, M., Furuya, T., Hashimoto, K., Oga, A., Oka, M. and Sasaki, K.: A copy number gain of the 6p arm is linked with advanced hepatocellular carcinoma. *J. Pathol.*, **9**：16, 2008.

5）Furuya, T., Uchiyma, T., Adachi, A., Okada, T., Nakao, M., Oga, A., Kawauchi, S., Kang, J.J., Yang, S-J. and Sasaki, K.: The development of a mini-array for estimating the disease states of gastric adenocarcinoma by array CGH. *BMC Cancer*, **8**：393, 2008.

6）Nakao, M., Kawauchi, S., Furuya, T., Uchiyama, T., Adachi, J., Okada, T., Ikemoto, K., Oga, A. and Sasaki, K.: Identification of DNA copy number aberrations associated with metastases of colorectal cancer using array CGH profiles. *Cancer Genet. Cytogenet.*, **181**：70-76, 2009.