Communications of the Korean Statistical Society Vol. 35, No. 3, 2008, pp. 265–271

Major SNP Marker Identification with MDR and CART Application

Jea-Young Lee[1], Yu-Mi Choi[2]

Abstract

It is commonly believed that diseases of human or economic traits of livestock are caused not by single genes acting alone, but multiple genes interacting with one another. This issue is difficult due to the limitations of parametric-statistic methods of gene effects. So we introduce multifactor-dimensionality reduction(MDR) as a methods for reducing the dimensionality of multilocus information. The MDR method is nonparametric (*i.e.*, no hypothesis about the value of a statistical parameter is made), model free (*i.e.*, it assumes no particular inheritance model) and is directly applicable to case-control studies. Application of the MDR method revealed the best model with an interaction effect between the SNPs, SNP1 and SNP3, while only one main effect of SNP1 was statistically significant for LMA ($p < 0.01$) under a general linear mixed model.

Keywords: Longissimus dorsi musle area(LMA); multifactor-dimensionality reduction(MDR); single nucleotide polymorphism(SNP).

1. Introduction

One traditional (logistic regression) approach to modeling the relationship between discrete predictors such as genotypes and discrete clinical outcomes (Hosmer and Lemeshow, 2000). However, logistic regression is limited in its ability to deal with interactions involving many factors. Having too many independent variables in relation to the number of observed outcome events is a well-recognized problem. In response to this limitation, Ritchie *et al.* (2001) developed the multifactor dimensionality reduction(MDR) approach that seeks to reduce the dimensionality of multilocus genotype space to facilitate the identification of gene-gene interactions. The MDR method is nonparametric (*i.e.*, no hypothesis about the value of a statistical parameter is made), model free (*i.e.*, it assumes no particular inheritance model) and is directly applicable to case-control studies (Ritchie *et al.*, 2001).

The data motivating this research were collected from Korean cattle (Lee *et al.*, 2008). The data was developed by the Hanwoo Improvement Center, National Agricultural

1) Professor, Department of Statistics, Yeungnam University, Kyungsan, 712-749, Korea. Correspondence: jlee@yu.ac.kr
2) Graduate Student, Department of Statistics, Yeungnam University, Kyungsan, 712-749, Korea. E-mail: chldbal84@hanmail.net

Cooperation Federation, Korea. The group composed of two hundred twenty nine steers from 16 grand-sire half-sibs families. The longissimus dorsi muscle area(LMA) were collected on all F1 progeny, were measured according to the standards of the Korean Animal Products Grading Service.

The QTL(quantitative trait locus) analysis for the daily gain and marbling scores in Korean cattle showed the ILSTS035 microsatellite as larger than the LOD score 3.0 (Kim *et al.*, 2003). Using statistical analysis has been to evaluate the association of three previously reported SNPs (12273_165(SNP1), 31465_446(SNP2), AH1_4(SNP3)) which were mapped at the same position with the ILSTS035 microsatellite in a linkage map of EST-based SNP (Snelling *et al.*, 2005) in BAT6, with the growth and carcass traits in Korean cattle (Lee *et al.*, 2008).

In this paper we search the statistical method for gene interaction effect analysis and identify major SNP interactions of LMA(longissimus dorsi muscle area) in Korean cattle.

2. Materials and Statistical Analysis

2.1. Materials

The data was developed by the Hanwoo Improvement Center, National Agricultural Cooperation Federation, Korea. The group composed of two hundred twenty nine steers from 16 grand-sire half-sibs families. The LMA data were collected on all F1 progeny and were measured according to the standards of the Korean Animal Products Grading Service.

2.2. MDR(multifactor-dimensionality reduction) method

The MDR method is model free (in that it does not assume any particular inheritance model) and is nonparametric (in that it does not estimate any parameters). Also, this method can be directly applied to case-control data and can easily search of multiple gene interaction (Bastione *et al.*, 2004). Consequently, the MDR analysis has been to search significantly effect (main effect and interaction effect) of the association between SNPs and LMA (economic traits) data. MDR method is divided into ten equal parts for data set (training set: 9/10 of the data, testing set: 1/10 of the data). Next, a set of n genetic factors is selected. The n factors are represented in n-dimensional space. During the model selection process, each multifactor cell class in n-dimensional space is labelled as high-risk if the ratio of cases to controls meets or exceeds the threshold of 1.0 and as low-risk otherwise. Among all of the two-factor combinations, the model that maximizes the cases to controls ratio is selected (This two-locus model has the minimum classification error among the two-locus models). To evaluate the predictive ability of the model, the model is evaluated of the 1/10 testing data set (prediction error). This procedure is done ten times. Cross-validation consistency(CVC) is a measure of the number of times (each trial) minimum ME model. Finally, maximum the cross-validation consistency and minimum PE are selected as the final best model.

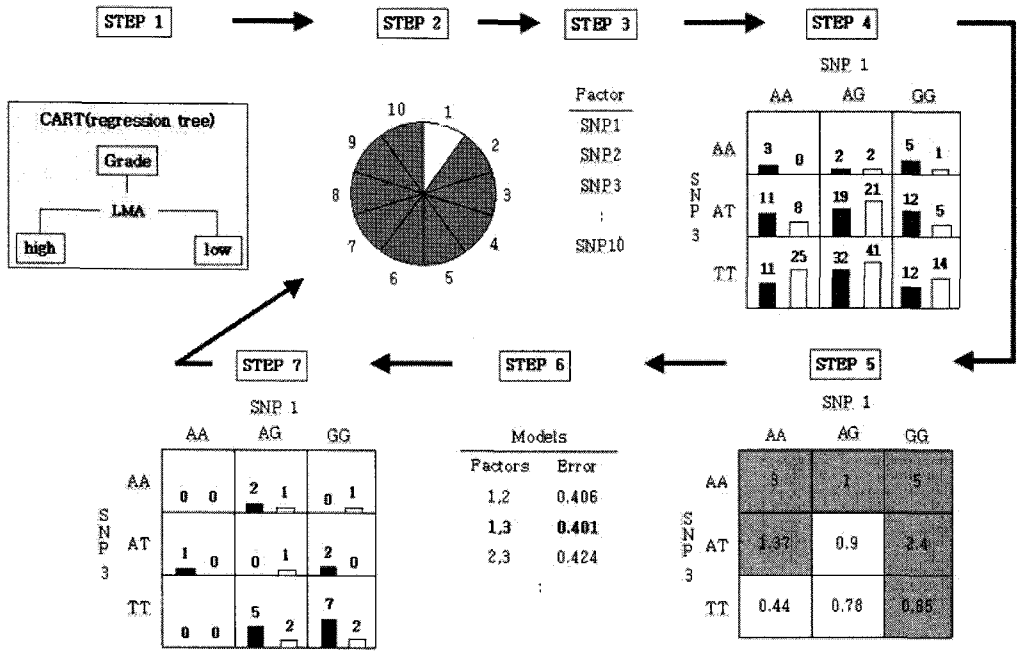

Figure 2.1: Summary of the general involved in implementation of the MDR method.

2.3. MDR application

Multifactor-dimensionality reduction(MDR) method is non-parametric and model-free and was initially implemented in case-control studies (Hahn *et al.*, 2003). For application to continuous data, the CART(classification and regression tree) algorithm was developed and combined into the MDR method. MDR method can detect the multiple gene interactions in case-control data only. Because LMA is not case-control data but continuous data, CART model is applied and is divided case-control data with MDR.

The MDR method is summarized in the following step-by-step procedures:

Step 1: LMA is divided into binomial data (case-control data) by CART model. In this initial implementation of MDR, case-control (1:1) data balanced.

Step 2: The data are randomly divided into 10 equal parts (training set: 9/10 of the data, testing set: 1/10 of the data).

Step 3: A set of n ($n = 1, 2$ or 3) genetic SNPs is selected from the pool of all SNPS (= 3).

Step 4: Based on the observed level of each of these SNPs, n factors and their possible multifactor classes or cells are represented in n-dimensional space.

Table 2.1: Average ME, average PE and CVC(10 repetitions) of SNPs.

Number of factors	Markers	average misclassification error(ME)	average prediction error(PE)	cross-validation consistency(CVC)
1	SNP1	0.42946	0.366	10
	SNP2	0.46070		0
	SNP3	0.43972	0.372	2
2	SNP1*SNP2	0.43124		0
	SNP1*SNP3	0.40758	0.348	8
	SNP2*SNP3	0.41204	0.352	2

Step 5: A cell is labeled as "high-ris" if the number of cases is equal to or greater than the number of controls and labeled as "low-risk" otherwise. For example, the cell in the first row and first column, so this cell is"high-ris".

Step 6: The MDR model with the smallest misclassification error (proportion of incorrect classification in training set, ME) is chosen among all of the two-factor combinations (*e.g.,* SNPs SNP1*SNP2, SNP1*SNP3, SNP2*SNP3).

Where, ME $= (Total_{high} - Case_{high}) + Case_{low}/N$

$Total_{high}$: the number of cases in high group

$Case_{high}$: the number of cases in low group

$Case_{low}$: the number of cases in low group

N: total number of data

Step 7: In order to evaluate the predictive ability of the model, the prediction error (proportion of incorrect prediction test set, PE) is estimated using 10-fold cross-validation.

Next, these seven steps were repeated for each possible combination of given n (SNP1, SNP2 and SNP3). Finally, maximum the cross-validation consistency and minimum average PE are selected as the final best model (Bastione *et al.*, 2004).

LMA (continuous data, $n = 229$) is divided into binomial data (case: 124, control: 105) by grade (using CART model), the balanced data (case: 124, control: 124) are obtained (using SPSS clementine 10.1). The result by MDR method represented in the Table 2.1.

Table 2.1 summarizes average ME, average PE and CVC for dierent combinations of SNPs that were obtained by applying the MDR method to LMA analysis. Among the models with single SNP, the marker SNP1 had the smallest average PE of 0.366. However, when considering interaction SNPs, the marker SNP1*SNP3 had the average PE of 0.348 which was smaller than that for the single SNP marker with SNP1 and thus the marker SNP1*SNP3 represented the best among all combinations of SNPs. Finally,

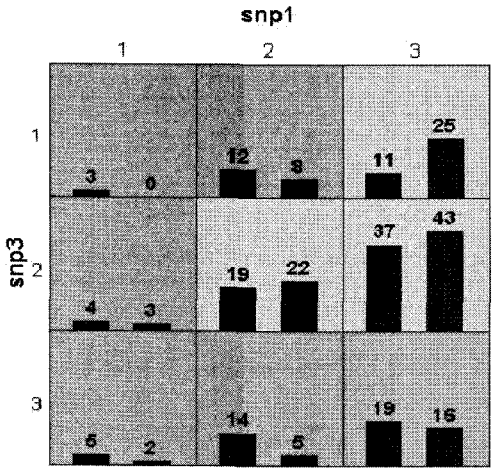

Figure 2.2: Distribution of high-risk and low-risk genotypes in the best two-locus model. (This distribution shows high-risk (dark shading) and low-risk (light shading) genotypes associated with economic traits(LMA) in the two-locus interaction detected by MDR analysis.)

the best-model marker SNP1*SNP3 was selected by evaluating the magnitude of cross-validation consistency and prediction error. But statistical signicance was not determined by PE and CVC. The permutation test will perform to determine emperical signicance thresholds by applying the same MDR model.

3. Permutation Test

Although the best-model was selected from MDR method, we could not determine statistical significance by MDR method. So statistical significances were derived from the permutation (10,000 times repeated) test. The null hypothesis that SNP1 (or SNP1*SNP3) does not affect significance the LMA, was rejected when the upper-tail Monte Carlo p-value derived from the permutation test was 0.05 or lower.

We will use the following procedures (Good, P., 2000) for a permutation test:

Step 1: Analyze the problem which is null hypothesis - SNP1 (or SNP1*SNP3) does not affect significance the LMA.

Step 2: LMA data are divided into binomial data (case-control data) by CART model. The data are divided into high, low group by MDR and compute sum of case (test statistics = 52) for high group.

Step 3: Compute sum obtained (sum of case) from the first sample (sum of case-control for high group, $n = 87$).

Table 3.1: p-value of permutation test for SNP1 and SNP1*SNP3.

Markers	permutation test p-value
SNP1	0.01618
SNP1*SNP3	0.00195

Step 4: Rearrange (permute) and repeat 10,000 times of step 3.

Step 5: Compare the new value of test statistic with the value you obtained from sum of case (test statistics = 52). Calculate significance probability using this permutation distribution of statistic.

Step 6: Step 2–5 are repeated 10 times. And compute average p-value. When emphp-value derived from the permutation test was 0.05, gene affect significance the LMA.

Table 3.1 is represented result of permutation test for SNP1 and SNP1*SNP3, obtained from 10,000 repetitions of LMA data set for each samples. SNP1 had a p-value of 0.01618 and SNP1*SNP3 had a p-value of 0.00195.

4. Results and Discussion

We have introduced MDR as a method for reducing the dimensionality of multiple gene information. It can detect the multiple gene interactions in case-control data. To detect the gene related to economic traits (continuous data) in Korean cattle, CART (classification and regression trees) model was divided case-control data and was applied with MDR. The results were obtained by applying the MDR method to LMA analysis. Among the models with single SNP, the model with SNP1 had the smallest average PE of 0.366. However, when considering interaction SNPs, the model with SNP1*SNP3 was selected and had the average PE of 0.348, which were lower than that for the single SNP model with SNP1. Thus interaction SNP1*SNP3 represented the best model among all combinations of SNPs.

Although the best-model was selected from MDR method, we could not determine statistical significance by MDR method. So statistical significances were derived from the permutation test. Permutation tests revealed that statistical significance was p-value ($p = 0.01618$) for SNP1 and $p = 0.00195$ for the interaction effects SNP1*SNP3. Consequently, the interaction gene (SNP1*SNP3) effect is more significant by the LMA than the single gene (SNP1) effect. Of course, MDR method can be computationally intensive and can be applied only to case-control studies that are balanced. Also, High dimensionality and a small sample lead to many multifactor cells with either missing data or singleton data. It is a problem for estimation of the prediction error. This limitation will be discussion in future studies.

References

Bastone, L., Reilly, M., Rader, D. J. and Foulkes, A. S. (2004). MDR and PRP: A comparison of methods for high-order genotype-phenotype associations. *Human Hered*, **58**, 82–92.

Good, P. (2000). *Permutation Test: A Ractical Guide to Resampling Method for Testing Hypotheses.* Springer-Verlag Berlin and Heidelberg GmbH & Co., New York.

Hahn, L. W., Ritchie, M. D. and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, **19**, 376–382.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons, New York.

Kim, J. W., Park, S. I. and Yeo, J. S. (2003). Linkage mapping and QTL on chromosome 6 in Hanwoo (Korean Cattle). *Asian-Australasian Journal of Animal Sciences*, **16**, 1402–1405.

Lee, Y. S., Bae, J. H., Lee, J. Y., Park, H. S. and Yeo, J. S. (2008). Identification of candidate SNP for economic traits on chromosome 6 in Korean cattle. *Animal Genetics*, submitting.

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F. and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics*, **69**, 138–147.

Snelling, W. M., Casas, E., Stone, R. T., Keele, J. W., Harhay, G. P., Bennett, G. L. and Smith, T. PL. (2005). Linkage mapping bovine EST-based SNP. *BMC Genomics*, **6**, 74–84.

[Received January 2008, Accepted February 2008]