BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 19 2007, pages 2589–2595
doi:10.1093/bioinformatics/btm396

Genetics and population analysis

Log-linear model-based multifactor dimensionality reduction method to detect gene–gene interactions

Seung Yeoun Lee[1], Yujin Chung[2], Robert C. Elston[3], Youngchul Kim[4] and Taesung Park[4,*]

[1]Department of Applied Mathematics, Sejong University, 98 Gunja-Dong Kwangjin-Gu, Seoul 143-747, Korea, [2]Department of Statistics, University of Wisconsin, 1300 University Avenue Madison, WI 53706, [3]Department of Epidemiology and Biostatistics, Case Western Reserve University, 10900 Euclid Avenue Cleveland, Ohio 44106-7281, USA and [4]Department of Statistics, Seoul National University, San 56-1 Shillim-Dong, Kwanak-Gu, Seoul 151-747, Korea

Received on March 7, 2007; revised on July 19, 2007; accepted on August 2, 2007

Associate Editor: Keith Crandald

ABSTRACT

Motivation: The identification and characterization of susceptibility genes that influence the risk of common and complex diseases remains a statistical and computational challenge in genetic association studies. This is partly because the effect of any single genetic variant for a common and complex disease may be dependent on other genetic variants (gene–gene interaction) and environmental factors (gene–environment interaction). To address this problem, the multifactor dimensionality reduction (MDR) method has been proposed by Ritchie *et al.* to detect gene–gene interactions or gene–environment interactions. The MDR method identifies polymorphism combinations associated with the common and complex multifactorial diseases by collapsing high-dimensional genetic factors into a single dimension. That is, the MDR method classifies the combination of multilocus genotypes into high-risk and low-risk groups based on a comparison of the ratios of the numbers of cases and controls. When a high-order interaction model is considered with multi-dimensional factors, however, there may be many sparse or empty cells in the contingency tables. The MDR method cannot classify an empty cell as high risk or low risk and leaves it as undetermined.

Results: In this article, we propose the log-linear model-based multifactor dimensionality reduction (LM MDR) method to improve the MDR in classifying sparse or empty cells. The LM MDR method estimates frequencies for empty cells from a parsimonious log-linear model so that they can be assigned to high-and low-risk groups. In addition, LM MDR includes MDR as a special case when the saturated log-linear model is fitted. Simulation studies show that the LM MDR method has greater power and smaller error rates than the MDR method. The LM MDR method is also compared with the MDR method using as an example sporadic Alzheimer's disease.

Contact: tspark@stats.snu.ac.kr

1 INTRODUCTION

In genetic epidemiology, it is a great challenge to identify and characterize susceptibility genes that have significant influences on the risk of common and complex diseases. This challenge is partly due to the limitations of parametric statistical methods for detecting genetic effects that are dependent on the degree of non-linearity in the relationship between genotype and disease. Non-linearities can occur from phenomena such as locus heterogeneity, phenocopies, and the dependence of genotypic effects on environmental factors (i.e. gene–environment interactions or plastic reaction norms) and genotypes at other loci (i.e. gene–gene interactions or epistasis) (Moore *et al*., 2006). From the statistical point of view, epistasis can be recognized as deviations from additivity in a linear statistical model. Epistasis is variously defined biologically and statistically. Biologically, epistasis is related to the physical interactions between biomolecules such as DNA, RNA and proteins, and occurs at the cellular level in an individual. On the other hand, statistical epistasis measures the non-additive effects of genes at the population level. It is difficult to detect and characterize epistasis because of non-linearity between genotype and disease (Moore *et al*., 2006). In the extreme case, epistasis might occur in the absence of detectable marginal effects of any one polymorphism.

Logistic regression is commonly used to model the relationship between genotypes and binary phenotypes. When high-order interactions involving multi-dimensional factors are considered, there may be many sparse or empty cells. In that case, the parameter estimates for the logistic regression model may have very large standard errors, resulting in an increase of type I error (Hosmer and Lemeshow, 2000). One solution for this problem is to collect very large samples to allow for robust estimation of interaction effects; however, the magnitudes of the samples that are often required may incur a prohibitive expense.

For relatively small sample sizes, Ritchie *et al*. (2001) proposed the MDR method to identify gene–gene and gene–environmental interactions with high-dimensional multilocus genotype variables for case-control and discordant-sib-pair studies. The MDR method collapses high-dimensional genotype variables into a one-dimensional variable with two levels (high and low risk) using the ratio of the number of cases to

*To whom correspondence should be addressed.

that of controls. The new, one-dimensional multilocus genotype variable is evaluated for its ability to classify and predict disease status through cross-validation. Many studies have shown that MDR can identify putative high-order gene–gene interactions in the absence of any significant independent main effects in sporadic breast cancer (Ritchie *et al*., 2001) and essential hypertension (Moore and William, 2002).

Hahn *et al*. (2003) developed a software package for implementing MDR in case-control and discordant sib-pair study designs. Ritchie *et al*. (2003) also evaluated the power of MDR in the presence of genotyping error, missing data, phenocopies and genetic heterogeneity. Furthermore, Coffey *et al*. (2004) compared MDR with the conditional logistic regression model for detecting gene–gene interactions on the risk of myocardial infarction, and pointed out the importance of model validation.

Although the MDR method provides many useful features, it has several drawbacks. First, when high-order interactions involving multi-dimensional factors are considered, there may be many sparse or empty cells in the contingency tables. In this case, the MDR method cannot classify the empty cells into high-risk and low-risk groups, which may result in loss of information and power for detecting gene–gene interactions. Second, MDR is prone to false positive and false negative errors when the number of cases is similar to the number of controls, or when the number of both cases and controls is too small. For example, suppose the ratio of cases to controls for a specific cell is equal to that for the entire set of cases and controls. Then, just a small change in this cell frequency can change the classification from the high-risk group to the low-risk group or vice versa. Thus, the classification of this cell is vulnerable to false positive and false negative errors.

In this article, we propose the log-linear model-based multifactor dimensionality reduction (LM MDR) method to reduce the loss of information due to the exclusion of empty cells, and to improve the validity of classification of MDR by using a log-linear model. The LM MDR method allows us to estimate the frequencies of empty cells as well as sparse cells, and classify them into the high-risk and low-risk groups. Moreover, since logistic models with categorical explanatory variables have equivalent log-linear models, forming the logit for one response variable helps interpret the results of the LM MDR method (Agresti, 2002). When the saturated model is fitted instead of a parsimonious model, LM MDR is equivalent to MDR, which implies that MDR is a special case of LM MDR. In practice, the non-saturated parsimonious models are preferable to the saturated one, since their fit smooths the sample data and have simpler interpretations (Agresti, 2002).

The main criticism of MDR is that it does not use a parsimonious model and allows biologically unlikely models. LM MDR is based on a parsimonious log-linear model and detects the main associations among variables in the model, which helps us find some obvious trends or patterns with biological plausibility.

The MDR method is briefly reviewed in Section 2.1 and the LM MDR method is proposed in Section 2.2. The LM MDR method is compared with the MDR method by simulation results in Section 3. The LM MDR method is compared with the MDR method in Section 4 using the data for sporadic Alzheimer's disease that was analyzed by Shi *et al*. (2005). Finally, a short discussion is given in Section 5.

2 METHODS

2.1 Multifactor dimensionality reduction method

The MDR method has been proposed by Ritchie *et al*. (2001) and Moore and William (2002), and implemented by Hahn *et al*. (2003) and Ritchie *et al*. (2003). It comprises the following two stages. Stage 1 involves choosing the best combination of multifactors. Stage 2 involves classifying the combinations of genotypes into high-risk and low-risk groups.

First, the data set is partitioned into 10 subsets for cross-validation. From those, 9 sets are assigned as a training set and the remaining 1 set taken as an independent test set. Second, for a combination of n given factors, all possible multilocus genotypes are represented in an n-dimensional contingency table. Each multilocus genotype in the n-dimensional space is then classified as 'high risk' if the ratio of the number of cases to the number of controls meets or exceeds some threshold, and 'low risk' if that threshold is not exceeded. Here, the threshold is determined as the ratio of the number of cases to the number of controls in the training set. This procedure reduces the n-dimensional space to a one dimensional space (i.e. one variable with two categories, high risk or low risk). Then, all the n-factor combinations are evaluated for their ability to classify the disease status in the training data set and the best combination of factors with the minimum misclassification error is selected. For the selected combination of factors, the prediction error is calculated in the test data set.

After repeating this procedure 10 times with the data split into 10 different training and test sets, the prediction error is averaged over the 10 data splits and is used as a measure of predictive power. In addition, a measure of the cross-validation consistency (CVC) is defined as the number of times a particular combination of factors is identified across the 10 cross-validations (Moore and William, 2002). These repeated procedures are performed 10 times, using different random numbers as seeds, to reduce the chance of observing spurious results due to chance divisions of the data.

For each value of n, we select the best combination of factors that has the minimum prediction error and the maximum CVC. Then, the best combinations can be listed, one for each possible dimension of n factors. That is, the result is a set of n models, one for each combination of factors. Then, from the selected combinations, the model with the combination of factors that maximizes the CVC and minimizes the prediction error is finally selected. When the CVC is maximal for one model and the prediction error is minimal for another model, the model with the smaller number of factors is selected.

Once MDR identifies the best combination of factors in Stage 1, the next step is to determine in Stage 2 which multilocus genotypes are high risk and which are low risk. MDR evaluates this final classification by a ratio threshold, the number of cases divided by the number of controls.

Recently, Moore *et al*. (2006) have described MDR as a constructive induction method in a flexible framework for detecting, characterizing and interpreting gene–gene interaction or epistasis. Thus, MDR creates a new attribute by pooling genotypes from multiple SNPs to capture interaction information. MDR is a constructive induction method that in its simplest form takes two or more variables and constructs a new variable, thereby changing the representation space of the data to make interactions easier to detect.

Since the first introduction of MDR, the accuracy (=1−error rate) has been used as a model fitness measure. It is calculated as the number of correctly classified samples divided by a total number of samples included in the evaluation. However, this measure is prone to produce biased results when the number of cases and controls are not the same.

To avoid this problem, Velez *et al.* (2007) suggested the measure balanced accuracy, defined as the arithmetic mean of sensitivity and specificity. When the number of cases and controls are the same, balanced accuracy is equal to accuracy.

Despite the recent improvement of MDR, it still suffers from the empty and sparseness problem. For example, MDR excludes empty cells from the analysis and leaves them as undetermined. In addition, the classification of MDR is vulnerable to errors when the number of cases is equal to the number of controls, or when both the numbers of cases and controls are too small. Since it is common to have empty or sparse cells in the contingency tables when high-order interactions involving multi-dimensional factors are considered, a remedy for taking into account the sparse and empty cells is needed in order to improve the validity of the MDR classification.

2.2 Log-linear model-based multifactor dimensionality reduction method

We propose the log-linear model-based multifactor dimensionality reduction (LM MDR) method to improve the validity of MDR in the classification procedure in Stage 2. The procedure in Stage 1 is the same as MDR in selecting the best combination of multifactors that has the minimum prediction error and the maximum CVC. In Stage 2, however, the LM MDR method classifies multilocus genotypes into the high-risk and low-risk groups using the estimated frequencies from the parsimonious log-linear model, whereas the MDR method uses the observed frequencies for its classification.

For LM MDR, the log-linear models are fitted for the best combination of factors selected in Stage 1. To find the parsimonious log-linear model, we compare all candidate log-linear models by the goodness-of-fit test statistic. For simplicity, we assume that the following three constraints are required to become candidate parsimonious models:

(i) The model is hierarchical.
(ii) The model has an equivalent logit model.
(iii) The model contains the interaction terms between the genotypes and the binary variable for distinguishing cases and controls.

By way of illustration, assume that the two SNPs each with three genotypes are selected as the best combination in Stage 1. Denote the two SNPs X and Y, and denote by D a binary variable distinguishing cases and controls. Then, the data for two SNPs and a binary variable are summarized in a $3 \times 3 \times 2$ contingency table. Let $\{\pi_{ijk}\}$, $\{\mu_{ijk}\}$ and $\{n_{ijk}\}$, for $i, j = 1, 2, 3$, and $k = 1, 2$, denote the cell probability, the expected cell frequency and the observed cell frequency, respectively. Then, the saturated log-linear model is defined by

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^D + \lambda_{ij}^{XY} + \lambda_{ik}^{XD} + \lambda_{jk}^{YD} + \lambda_{ijk}^{XYD},$$

where λ_i^X, λ_j^Y and λ_k^D are the main effects of X, Y and D, respectively; λ_{ij}^{XY} represents the two-way interaction effect of X and Y, and analogously for λ_{ik}^{XD} and λ_{jk}^{YD}; and λ_{ijk}^{XYD} represents the three-way interaction effect of X, Y and D. This saturated model is represented by a symbol, (XYD), that lists the highest-order term(s). The maximum-likelihood estimators (MLEs) of μ_{ijk} are given by $\hat{\mu}_{ijk} = n_{ijk}$ for all i, j and k. That is, the MLEs are just the observed cell frequencies.

The list of all possible log-linear models includes (XYD), (XD, YD, XY), (XD, XY), (YD, XY), (XD, YD), (X, YD), (XD, Y), (D, XY) and (X, Y, D). However, the models (XYD) and (XD, YD, XY) are the only candidate log-linear models satisfying the above three constraints. The model (XD, YD, XY) is defined by

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^D + \lambda_{ij}^{XY} + \lambda_{ik}^{XD} + \lambda_{jk}^{YD}.$$

For this log-linear model, the MLEs of $\hat{\mu}_{ijk}$ are obtained by iterative methods and the estimated cell frequencies are not equal to the observed cell frequencies.

The adequacy of the log-linear models can be tested by the goodness-of-fit test statistic. If there are several candidate models available with good fit to the data, then that model with the least number of parameters is selected. Alternatively, Akaike's information criterion (AIC) can be used and the model with the smallest AIC value is selected as the best model among the several candidate models.

The estimated cell frequencies from the selected log-linear model can be used to classify the multilocus genotypes into high or low-risk groups depending on the ratio of the estimated number of cases to the number of controls, i.e. $\hat{\mu}_{ij1}/\hat{\mu}_{ij2}$. The threshold ratio is the ratio of the estimated number of cases to the estimated number of controls in the entire data set. The selected log-linear model provides the MLEs of the frequencies of both empty cells and sparse cells, and thus allows these cells to be classified into either the high or low-risk group. Moreover, since the selected log-linear model accounts for the significant association between susceptible genotypes and the disease, the estimated frequencies from the selected model become quite informative. As a result, the classification of LM MDR is expected to be more accurate than that of MDR.

When no candidate models are selected with a good fit to the data, the saturated model is selected. The estimated cell frequencies are then the same as the observed frequencies, i.e. $\hat{\mu}_{ijk} = n_{ijk}$ for all i, j and k. Thus, in this case the LM MDR method yields the same result as that of the MDR method. In other words, when the saturated log-linear model is used, the MDR method becomes a special case of the LM MDR method.

3 SIMULATION

We perform a simulation study to compare the LM MDR method with the MDR method in terms of power, misclassification and prediction errors. We assume a case-control study using similar two-locus epistasis models considered by Ritchie *et al.* (2003) and Velez *et al.* (2007). However, our models differ from theirs in that ours are based on the log-linear model. Our simulation settings extend from a model with interaction effects in the absence of marginal effects to a model with only marginal effects in the absence of interaction effects.

One set of the multilocus penetrance functions assumed is given in Table 1. This set has some main effects as well as two-way interaction effects. Here penetrance is defined by $P[D = 1 \mid X = i, Y = j]$. We assume the major allele frequencies of X and Y are each 0.8, and joint Hardy-Weinberg equilibrium at the two loci. Then, the probabilities of the genotypes are given by

$$\begin{aligned} P[X = 1, Y = 1] &= 0.4096, \\ P[X = 2, Y = 2] &= 0.1024, \\ P[X = 3, Y = 3] &= 0.0016, \\ P[X = 1, Y = 2] = P[X = 2, Y = 1] &= 0.2048, \\ P[X = 1, Y = 3] = P[X = 3, Y = 1] &= 0.0256, \\ P[X = 2, Y = 3] = P[X = 3, Y = 2] &= 0.0128. \end{aligned}$$

Here, the values of 1, 2 and 3 represent the three different genotypes, e.g. AA, Aa and aa, respectively. Using $P[X, Y]$ and the penetrance functions given in Table 1, we generated 100 different data sets each with 100 cases and 100 controls. The true high-risk and low-risk groups are shown in Table 2.

Table 1. Multilocus penetrance function

Genotypes	$Y=1$	$Y=2$	$Y=3$
$X=1$	0.1514	0.0002	0.0232
$X=2$	0.1033	0.0001	0.0151
$X=3$	0.5959	0.0021	0.1647

Table 2. The true high-risk (=1) and low-risk (=0) simulated

Genotypes	$Y=1$	$Y=2$	$Y=3$
$X=1$	1	0	0
$X=2$	1	0	0
$X=3$	1	0	1

Since the LM MDR method has the same selection procedure as the MDR method, we assume that the best combination of SNPs is (X, Y) in Stage 1. For the LM MDR method, a model, (XD, YD, XY), is fitted to each of the 100 data sets. The average values of the 100 goodness-of-fit test statistics and P-values are 1.15 and 0.86 with $df=4$, respectively, showing that this model fits the 100 data sets well.

We compare three methods, which are LM MDR without continuity correction, LM MDR with continuity correction and MDR. As shown in Table 3, the two LM MDRs have slightly smaller prediction errors than MDR, but similar misclassification errors to MDR. However, note that the MDR method excluded the empty cells, while the two LM MDR methods included them in computing the misclassification and prediction errors.

The three methods are also compared in terms of two different measures: the number of perfectly matched data sets and the percentage of matched cells. The number of perfectly matched data sets represents how many times 100 data sets have perfectly matched the true high-risk and low-risk groups. The percentage of matched cells represents the average percentage of the nine cells that match both the true high and low-risk groups.

As shown in Table 3, the LM MDR methods without and with continuity correction yielded 36 and 58 perfectly matched sets, respectively, while the MDR method yielded 29 matched sets. That is, these two LM MDR methods classified more perfectly than did MDR, and LM MDR with continuity correction classified all nine cells perfectly 22 times more often than LM MDR without continuity correction. Furthermore, LM MDR methods yielded a higher percentage of individual matching cells than MDR.

We have presented in detail just one typical result from among the many simulation studies we performed with two loci in which several combinations of main effects and interaction effects were considered. All these simulation, the results of which we do not detail here, agreed qualitatively with what might be expected from theoretical considerations, as we summarize in the discussion. If an unsaturated model provides a good fit to the data (defined in these simulations as no departure from the model at the 5% significance level), then the LM MDR methods have better power and smaller prediction errors than the MDR method, and the classification of the LM MDR method can be substantially improved by using the continuity correction.

Table 3. Comparison of results between three methods

Methods	MDR	LM MDR[1]	LM MDR[2]
Classification error	0.2960	0.2967	0.3052
Prediction error	0.3137	0.3091	0.3065
Number of perfect matching data sets[3]	29	36	58
Percentage of matching cells[4]	86.22	90.33	94.56

[1]LM MDR without continuity correction.
[2]LM MDR with a continuity correction.
[3]The number of data sets among the 100 data sets that perfectly matched the true high-and low-risk groups.
[4]The average percentage of cells that matched the true high- or low-risk groups.

4 EXAMPLE

Alzheimer's disease, the most common form of dementia in elderly persons, is a progressive neurodegenerative disorder. Deposition of amyloid β-peptide (Aβ) in brain from Alzheimer's disease patients is one of the pathological hallmarks of Alzheimer's disease, and neprilysin (NEP) is a major β-amyloid peptide (Aβ)-degrading enzyme *in vivo* and reduced mRNA levels of NEP correlates with increased plaque density. The human NEP gene is composed of 24 exons located on chromosome 3q25.1–q25.2 (D'Adamio *et al*., 1989) and familial Alzheimer's disease is reported to be linked to 3q23–q24 (Poduslo *et al*., 1999).

Recent studies have focused on enzymes in amyloid catabolism. NEP, a putative Aβ degrading enzyme, is reported to be important in the development of Alzheimer's disease since its decreased expression and/or activity may also result in cerebral Aβ accumulation. Helisalmi *et al*. (2004) investigated whether allelic variants across the NEP gene modify the risk of Alzheimer's disease in a Finnish population and found that two SNPs, rs989692 and rs3736187 in the NEP gene affect the risk of Alzheimer's disease in a study of 390 Alzheimer's Finnish patients and 468 cognitively healthy controls.

Shi *et al*. (2005) investigated the association of the NEP gene on the development of Alzheimer's disease in a Chinese sample that consisted of 257 Alzheimer's disease patients and 242 age-matched controls. They used denaturing high-performance liquid chromatography (DHPLC) to screen the NEP gene for SNPs and then each candidate SNP was confirmed by DNA sequencing. As a result, they identified eight novel and one known SNP and found that $-204G\rightarrow C$ in the promoter region, and IVS17$-294C\rightarrow T$, and IVS22$+36C\rightarrow A$ in an intron, show a significant association with Alzheimer's disease. In addition, the subsequent haplotype

Table 4. Result of stage 1

SNPs in the best combination in each model	Prediction error	CV consistency
IVS22+36C→ A, 3′UTR159C→T,	0.4510	**9**
−204G → C, IVS22+36C→ A, 3′UTR159C → T	**0.4177**	**9**
IVS10-5C → T, IVS15+144T → A, IVS22+36C→ A, 3′UTR159C → T	0.4254	7

The maximum cross-validation consistency and minimum predicition error is indicated in bold type.

(a)

Y \ X	CC	CA	AA
CC	124:102	8:2	0:1
CT	73:99	30:11	0:0
TT	16:22	5:5	1:0

(b)

Y \ X	CC	CA	AA
CC	124:102	8:2	0:1
CT	73:99	30:11	0:0
TT	16:22	5:5	1:0

Fig. 1. (**a**) LM MDR method (**b**) MDR method. Distribution of high-risk and low-risk genotypes for the best two-locus model, where X denote a categorical variable representing the genotypes for SNP of IVS22+36C → A and Y for SNP of 3′UTR159C → T. This summary of the distribution shows high-risk (dark shading), low-risk (light shading) and undetermined risk (empty cell, white) genotypes provided by the LM MDR method and MDR method. The numbers in the cell represent the number of case (left) and control (right).

analysis involving these three SNPs further confirmed a significant association with Alzheimer's disease.

In this article, we applied the MDR and the LM MDR methods to the data from Shi *et al*. (2005) with all possible combinations of the eight SNPs up to the fourth order. One of the nine SNPs was excluded from the analysis owing to its rare allele frequency (<0.01). Table 4 summarizes the CVC and the prediction errors obtained from the MDR method.

As shown in Table 4, the best combination set for the two-locus models has a maximum CVC of 9 with a prediction error of 0.4510, whereas the best combination set for the three-locus models has a minimum prediction error of 0.4177 with the CVC of 9. Thus the three-locus model should be selected as the best MDR model since it has the maximum CVC as well as the minimum prediction error. However, since the difference of the prediction errors between these two models is very small, the two-locus model was selected as the best MDR model to investigate the association of SNPs on Alzheimer's disease. In fact, when the three-locus model is considered, more than half of the cells are empty and excluded from the classification procedure by MDR. Furthermore, the fitted log-linear model for the LM MDR method yields no meaningful interaction effects among the three SNPs.

To classify the two-locus genotypes into the high-risk and low-risk groups in Stage 2, the MDR method uses the observed frequencies in a $3 \times 3 \times 2$ contingency table, while the LM MDR method uses the estimated frequencies from the parsimonious log-linear model. For the LM MDR method, we fit the model (XD, YD, XY), where D denotes a binary response variable representing Alzheimer's disease and control, X denotes a categorical variable representing the genotypes for SNP IVS22+36C → A, and Y for SNP 3′UTR159C → T, respectively. The goodness-of-fit test statistic is 4.97 with $p>0.1$ ($df=4$), which implies that this model fits the data well. Since this model is the only candidate model satisfying the three constraints, this log-linear model is used for classifying the two-locus genotypes.

The two tables in Figure 1 show how the two genotype combinations are classified into the high-risk and low-risk groups by MDR and LM MDR, respectively. Here note that the two LM MDR methods yielded the same classification results regardless of using the continuity correction or not. As a result, the LM MDR method classified five genotype combinations into the high-risk group and four genotype combinations into the low-risk group (Fig. 1a), whereas the MDR method classified four into the high-risk group and four into the low-risk groups with one empty cell undetermined (Fig. 1b). The MDR classification procedure excluded the empty cell, $(X,Y)=(AA,CT)$, while the LM MDR method classified it into the low-risk group. Comparing MDR with LM MDR, there are three inconsistent results for the cell $(X,Y)=\{(CA,TT), (AA,CC), (AA,TT)\}$. These cells have equal frequencies of cases and controls, (5:5), or too small frequencies of case and controls, (0:1) and (1:0). Among those, (5:5) and (0:1) are classified into the low-risk groups by MDR but into the high-risk groups by LM MDR, whereas (1:0) is classified into the high-risk group by MDR but into the low-risk group by LM MDR. This implies that the classification procedure is vulnerable to false positive or false negative errors when the numbers of cases and controls are equal, or too small. The two methods, however, yielded consistent classification results when there was a large difference in the numbers of cases and controls in the cells.

From the results of the LM MDR method, it seems that the combination of IVS22+36C/A and 3′UTR159C/C has significant association with increased risk of Alzheimer's disease while the MDR method yields a rather inconsistent trend. According to the results of Shi *et al*. (2005), IVS22+36C/A is one of the significant SNPs associated with Alzheimer's disease. However, it was reported that IVS22+36C/A seemed unlikely to influence

neprilysin expression levels or activity, because it is located in an intron and is 36 bp and ~ 300 bp away from the flanking splice sites. Instead, it was proposed that IVS22+36C/A is in linkage disequilibrium with other 'true' Alzheimer's disease risk variants within or near the NEP gene. For example, c.401A → G might be such a candidate, although it has no significant association with Alzheimer's disease in this study. On the other hand, 3′UTR159C/C was found by Clarimon *et al*. (2003) to be associated with Alzheimer's disease in persons less than 75 years old in a Spanish population.

In summary, it has been shown that two SNPs, IVS22+36C → A and 3′UTR159C → T in the introns, are interactively associated with Alzheimer's disease. In particular, the combinations of IVS22+36A and 3′UTR159C are significantly associated with increased risk of Alzheimer's disease. These results are also in line with those of the haplotype analysis (Shi *et al*., 2005). As a result, comparing the LM MDR method with the MDR method, LM MDR provides more consistent results than MDR as well as including all empty cells in the classification procedure, while MDR excludes them.

5 DISCUSSION

In this article, we have proposed the LM MDR method to improve MDR in classifying sparse or empty cells. LM MDR uses the log-linear model to estimate the cell frequencies for the best combination of factors selected in Stage 1. The LM MDR method includes the MDR method as a special case when the saturated model is selected.

We performed many simulation studies with two loci in which several combinations of main effects and interaction effects were considered. All these simulation agreed qualitatively with what might be expected from the following theoretical considerations. In general, the unsaturated log-linear model provides the maximum-likelihood (ML) estimates that smooth the sample frequencies. The unsaturated log-linear model can result in a smaller mean square error (MSE), defined as the sum of the variance and the squared bias, than the saturated model. Although they may be biased, the estimates have smaller variances because they are based on estimating fewer parameters than are required for the saturated model, unless the sample size is so large that the bias term dominates the variance. A simple illustration is given in Agresti (2002).

It is well known that empty cells and sparse tables can cause problems with the existence of estimates for log-linear model parameters, problems with severe bias of odds ratios, problems with the performance of computational algorithms, as well as problems with asymptotic approximations of chi-squared statistics. The ML estimate for the empty cell is equal to zero for the saturated model since the ML estimates are equal to the observed frequencies for the saturated model. On the other hand, the unsaturated models tend to provide non-zero ML estimates (Agresti, 2002). The MSEs of the empty cells or sparse cells can be computed for both saturated and unsaturated models. In most cases, the unsaturated model provides smaller MSEs for these cells than the saturated model.

As expected, the extensive simulation studies showed that the LM MDR has less prediction errors than MDR when there are some marginal effects. However, LM MDR tends to have similar prediction errors to MDR as the marginal effects decrease. In this case, the unsaturated model does not fit the data well. Thus, LM MDR uses the saturated model, which yields the same result as MDR. However, MDR and LM MDR do not show much difference in misclassification errors and prediction errors when the interaction effects are small. This reflects the importance of the presence of marginal effects on fitting the log-linear model and the importance of including in the LM MDR procedure a goodness-of-fit test to obtain the best parsimonious model.

We also compared the power, in which power is defined as either the number of perfectly matched data sets or percentage of matched individual cells. LM MDR has larger power than MDR in terms of the number of perfectly matched data sets, when there are some marginal effects, regardless of the presence of interaction effects. Further, LM MDR has larger power than MDR in terms of percentage of matched individual cells, when there are some marginal effects. The reason for this is that MDR excludes empty cells but LM MDR classifies empty cells by using the estimated frequencies from the unsaturated log-linear model. LM MDR has less power than MDR only when there are no marginal effects. These simulation results imply that LM MDR is better than MDR for a model in which there are even small marginal effects, whereas MDR is better when there are only higher-order interaction effects in the absence of marginal effects, and then only provided the individual sample cell frequencies are large enough.

In summary, when there is high-order epistasis in the absence of marginal effects, the LM MDR procedure that includes finding a parsimonious model provides a similar result to that of the original MDR approach. Otherwise, LM MDR provides a better result than the original MDR approach, because one of the strengths of the unsaturated log-linear model is to estimate the cell frequencies in the sparse or empty cells when the unsaturated model gives a good fit to the data.

Both MDR and LM MDR are model free in the sense that no particular genetic model is assumed, and are non-parametric in the sense that no genetic model parameters are estimated (Ritchie *et al*., 2001). However, it is easier to interpret the result of the LM MDR model than that of the MDR model, because log-linear models have equivalent logit models that are useful in describing obvious trends or patterns in terms of odds ratios. In addition, empty cell and sparseness problems become more serious the larger the number of loci in the model, so that when high-order interactions with multifactors are considered, LM MDR is more informative than MDR: LM MDR provides the estimated frequencies of the empty cells using the log-linear model while MDR excludes the empty cells from the analysis.

For further study, we are considering extending the LM MDR to use the log-linear model in selecting the best combination of genotypes in Stage 1. The best combination of factors is selected by evaluating the prediction errors and CVC among all possible combinations of factors. Since the prediction errors and CVC depend on how correctly each cell is classified, it may be more informative to use the log-linear model in this stage as well. This may end up with a biologically more meaningful model to reflect the relationship between the disease and multifactors.

The higher-order model involving three of four SNPs includes a more complicated process for fitting the unsaturated model and we are now developing a more efficient way of model-fitting. Typically, the number of empty cells increases very rapidly with the number of loci, because the model then introduces a contingency table with many cells. We anticipate that the LM MDR will handle this problem appropriately.

When applying the MDR and LM MDR methods, the presence of missing observations reduces the number of observations available in the analysis. Although an extension of MDR has been proposed for handling missing observations, it is still at an early stage of development. The most appropriate approach at present is to use only subjects with complete observations. However, as the number of genotypes increases, the number of subjects with complete observations decreases rapidly, because any subject with at least one missing observation for any genotype has incomplete observations. One solution to handle this situation is to impute missing observations. Imputation of missing observations can reduce empty or sparse observations. We will combine the imputation of missing observations and the LM MDR method in a future study.

ACKNOWLEDGEMENTS

The work was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126) and a US Public Health Service Research grant (GM28356) from the National Institute of General Medical Sciences.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (2002) *Categorical Data Analysis*, 2nd edn. John Wiley and Sons, New Jersey, pp. 85–86.

Clarimon,J. *et al*. (2003) Possible increased risk for Alzheimer's disease associated with neprilysin gene. *J. Neural Transm.* **110**, 651–657.

Coffey,C.S. *et al*. (2004) An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene–gene interactions on risk of myocardial infrcation: the importance of model validation. *BMC Bioinformatics*, **5**, 49.

D'Adamio,L. *et al*. (1989) Organization of the gene encoding common acute lymphoblastic leukemia antigen (neutral endopeptidase 24.11):multiple miniexons and separate 5′ untranslated regions. *Proc. Natl. Acad. Sci. USA*, **86**, 7103–7107.

Hahn,L.W. *et al*. (2003) Multifactor-dimensionality Reduction software for detecting gene–gene and gene-environment interactions. *Bioinformatics*, **19**, 376–382.

Helisalmi,S. *et al*. (2004) Polymorphisms in neprilysin gene affect the risk of Alzheimer's disease in Finnish patients. *J. Neurol. Neurosurg. Psychiatry*., **75**, 1746–1748.

Hosmer,D.W. and Lemeshow,S. (2000) *Applied Logistic Regression*. 2nd edn. John Wiley and Sons, New York.

Moore,J.H. and William,S.M. (2002) New strategies for identifying gene–gene interactions in hypertension. *Ann. Med.*, **34**, 88–95.

Moore,J.H. *et al*. (2006) A flexible computational framework for detecting, characterizing, and interpreting statistical pattern of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.*, **241**, 252–261.

Poduslo,S.E. *et al*. (1999) A familial case of Alzheimer's disease without tau pathology may be linked with chromosome 3 markers. *Hum. Genet*., **105**, 32–37.

Ritchie,MD. *et al*. (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet*., **69**, 138–147.

Ritchie,M.D. *et al*. (2003) Power of Multifactor-dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Gene. Epidemiol.*, **24**, 150–157.

Shi,J. *et al*. (2005) Mutation screening and association study of the neprilysin gene in sporadic Alzheimer's disease in Chinese persons. *J. Gerontol. Biol. Sci*., **60A**, 301–306.

Velez,D.R. *et al*. (2007) A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Gene. Epidemiol*, **31**, 306–315.