응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 53–63*

# Multifactor Dimensionality Reduction(MDR)을 이용한 한우 도체중에서의 주요 SNP 규명*

이제영[1] 김동철[2]

요 약

일반적으로 인간의 질병과 가축의 경제적인 특성은 하나의 유전자가 아닌 여러 유전자의 상호작용으로 일어난다고 믿고 있다. 따라서 본 연구에서는 세대를 거듭할수록 대립유전자의 유전이 안정적으로 발생되어지고 개체의 기능적인 유전적 가치를 직접적으로 추정할 수 있는 single nucleotide polymorphism(SNP)을 한우의 경제적 특성인 도체중(carcass cold weight)에 대하여 모수적인 방법인 ANOVA와 비모수적인 방법인 multifactor dimensionality reduction(MDR)을 이용하여 하나의 유전자의 효과와 두 개의 유전자의 상호작용 효과를 비교하였다. ANOVA에서는 하나의 유전자 SNP1이 도체중에 유의한 효과가 있었고 상호작용 효과에서는 도체중에 유의한 효과는 없었다. MDR에서는 하나의 유전자의 효과인 SNP1과 두 개의 유전자의 상호작용인 SNP1*SNP2의 효과가 컸으며 SNP1과 SNP1*SNP2를 비교했을 시에는 SNP1*SNP2의 효과가 더 크게 나타났다. 이는 개별 SNP유전자 보다 복합 SNP유전자의 상호작용이 경제적인 특성인 도체중에 더 영향을 준다는 것을 알 수 있었다.

주요용어: SNP, ANOVA, MDR, 도체중, CART(classification and regression trees).

## 1. 서론

인간의 질병 또는 가축의 경제적인 특성에 관한 유전자의 규명은 유전학에서 매우 중요한 관심사이다. 일반적으로 인간의 질병과 가축의 경제적인 특성은 하나의 유전자가 아닌 여러 유전자의 상호작용으로 일어난다고 믿고 있다. 그래서 이런 여러 유전자의 상호작용을 고려한 모형으로 선형모형 같은 표준 통계적 모델로 사용해왔다. 그러나 유전자의 수가 많을 경우 상호작용의 조합이 많아지므로 종종 모수들의 상호작용에 대한 해석과 모형을 결정하는 것이 어려울 수 있다. 그리고 비록 여러 유전자의 상호작용에 대하여 모형화된 경우라도 많은 가능한 테이블 셀에 관측값이 없을 수도 있다. 이 경우에 추정값과 오차가 더 크게 나타날 수 있다 (Hosmer와 Lemeshow, 2000). 그래서 우리는 여러 유전자에 대한 상호작용을 결정하는 방법으로 Multifactor Dimensionality Reduction을 소개한다(Ritchie 등,

* 본 연구는 농림부 선정 경북한우사업단의 지원(2006)으로 수행되었음.
1) (712-749) 교신저자. 경북 경산시 대동 214-1 영남대학교 통계학과, 교수.
E-mail: jlee@yu.ac.kr
2) (712-749) 경북 경산시 대동 214-1 영남대학교 통계학과, 대학원, 석사.
E-mail: kitty2142@ynu.ac.kr

2001). Multifactor Dimensionality Reduction(MDR)은 상호작용에 대한 명확한 모형의 가정 없이 비모수적인 방법으로 적당한 high-order 차수의 데이터로 복잡한 관계를 밝힐 수 있다. 그래서 본 연구에서는 한우의 경제형질인 도체중(carcass cold weight)에 관련된 주요 유전인자를 MDR 방법을 이용하여 찾아보려고 시도 하였다.

본 연구의 데이터는 Korean cattle (Lee 등, 2007)에서 수집되었으며 농협중앙회 가축개량 사무소에서 개발되었고 16 grand-sire half-sibs families로부터 229두의 수송아지로 구성되어졌다. 도체중은 모든 F1 자손으로부터 수집되어졌고 한국축산물등급판정소의 규격에 따라 측정 되었다. 대부분의 QTL 연구는 집단의 규모가 크거나 반형매(half-sib)에서 sire의 대립유전자가 유전되는 microsatellite marker를 사용한다. 그러나 QTL 연구의 결과들을 현장에 적용하기가 극히 제한적이다. 왜냐하면 기준집단에서 QTL은 다음세대로 유전되어지기 어렵기 때문이다. 따라서 많은 세대를 거듭할수록 대립유전자의 유전이 안정적으로 발생되어지고 개체의 기능적인 유전적 가치를 직접적으로 추정할 수 있는 single nucleotide polymorphism(SNP)을 한우의 경제적 특성에 대하여 genetic test의 개발에 이용한다.

현재까지 소에서는 도체형질(도체중량, 등지방두께, 등심단면적, 근내지방도)과 연관이 있는 SNP marker들이 일반가축에서 평가되어지거나 적용되고 있다 (Barendse 등, 2004; Page 등, 2004). 따라서 본 연구에서는 EST-based SNP 연관지도 (Snelling 등, 2005)에서 Kim 등 (2003)에 의해 규명 되어진 한우 염색체 6번에 위치한 candidate QTL인 IL-STS035 microsatellite marker 와 같은 거리에 있는 SNP들 중 polymorphisms가 나타난 31465_446(SNP1), 12273_165(SNP2), AH1-4(SNP3)를 선발 (Lee 등, 2007)하여 국가 후대검정우 229두에 대한 경제형질의 도체중에 대하여 ANOVA(Analysis of variance)선형 모형으로 분석하였다. 그러나 앞에서 말한 것과 같이 상호작용의 모형은 모수의 상호작용에 대한 해석과 모형의 결정이 어려울 수 있다. 그래서 상호작용의 복잡한 관계를 밝힐 수 있는 MDR 방법을 적용해 보려고 한다. 이 연구에서 데이터의 경제형질의 도체중이 연속형 자료이기 때문에 먼저 데이터마이닝 기법중 하나인 CART(classification and regression trees)방법으로 case-control로 이분화 시킨 후에 MDR 방법에 적용시켰다. 본 논문은 다음과 같이 구성되었다. 2절에서는 앞에 사용한 분석방법인 MDR에 대하여 소개를 하고 3절에서는 ANOVA, MDR을 적용한 결과를 보여준다. 4절에서는 연구결과를 요약 정리하였다.

## 2. Multifactor-Dimensionality Reduction(MDR) 분석 방법

MDR 방법은 일반화된 선형 모형인 전통적인 통계 기법과는 달리 어떤 모수에 대한 추정과 genetic 모형의 가정을 요구하지 않는다(다시 말하면 특별한 유전형질 모형에 대한 가정이 필요 없다). Ritchie 등 (2003)과 Hahn 등 (2003)에 의하면 MDR 모형에서 처음으로 시행하는 것은 case와 control을 1:1로 균형을 맞추는 것이다. 다음 step에서 case-control에 대한 MDR 방법을 시행하는 과정을 보여준다.

**Step 1.** 데이터를 랜덤으로 10개의 같은 크기로 나눈다. 그리고 그중 9개를 training set 으로 1개를 testing set으로 둔다.

**Step 2.** 모든 SNP로부터 $k$개의 SNP조합 중 하나를 선택한다.

**Step 3.** 선택된 SNP조합에서 SNP의 각각 수준을 기초로 한 개체들을 multifactor classes 또는 cells에 기술한다. 예를 들어서 $k = 2$ 일 경우 SNP는 3개의 수준으로 되어있다. 따라서 9개의 셀을 가진다. 각각 9개의 셀에 case의 값과 control의 값을 적는다.

**Step 4.** Case와 control의 비를 구하여 1보다 크거나 같으면 high-risk, 1보다 작으면 low-risk라 한다. 예를 들면 1행 1열의 경우 case와 control의 비가 1보다 작으면 이 셀은 low-risk이다.

**Step 5.** K개의 SNP의 조합 전부에서 데이터의 9/10인 traing set에서 잘못 분류된 비율인 misclassification error(ME)를 구한다. 여기서 *잘못 분류된 비율인 ME*는 $\{Total_{high} - Case_{high} + Case_{low}\}/N$ 이다. $Total_{high}$ 는 high그룹의 전체 값이며 $Case_{high}$ 는 high그룹의 전체 case 경우의 수이며 $Case_{low}$는 low그룹의 전체 case 경우의 수이다. 그리고 N은 전체 데이터의 수이다. 이렇게 구한 ME들 중에 가장 작은 값을 선택한다.

**Step 6.** Training set에서 high-low로 나눈 표를 나머지 1/10의 데이터인 testing set을 이용하여 *잘못 분류된 비율인 prediction error(PE)*를 Step 5의 정의와 같이 구한다.

그 다음 위의 과정의 반복에서 나온 10개의 ME와 PE의 평균을 구해 그 값이 가장 낮은 것을 best $n$-factors 모형으로 정한다 (Bastione 등, 2004). 그리고 앞에서 구한 각각의 ME를 이용하여 cross validation consistency(CVC)를 구하는데 이것은 10번의 cross- validation을 시행할 때 각 시행에서 선택된 best model을 카운트하는 것이다(Chung 등, 2005). 따라서 ME와 PE의 평균이 가장 낮고 CVC가 가장 높은 값이 best n-factors 모형이다.

## 3. 실험자료와 결과

### 3.1. 실험 자료

이 데이터는 농협중앙회 가축개량 사무소에서 개발되었고 16 grand-sire half-sibs families로부터 229두의 수송아지로 구성되어졌다. 도체중은 모든 F1 자손으로부터 수집되어졌고 한국축산물등급판정소의 규격에 따라 측정 되었다. 그리고 polymorphisms가 나타난 12273_165(SNP1), 31465_446(SNP2), AH1-4(SNP3)를 이용 하였다 (Lee 등, 2007).

### 3.2. ANOVA를 이용한 도체중에 대한 SNP 영향

한우의 양적형질에 대한 유전분석에 있어 가장 기본이 되는 개념은 어느 개체의 표현형이 그 개체의 유전자형에 의한 효과와 환경의 효과에 의하여 결정된다는 것이다. 즉, 개체에 대한 표현형은 다음과 같이 두 부분으로 나눌 수 있다.

$$P = G + E. \tag{3.1}$$

여기서 P는 표현형이고 G는 유전자효과, E는 환경효과라 한다. 그래서 가축의 경제형질 연구에서는 일반적으로 다음과 같은 통계적 모델을 사용한다.

표 3.1: SNP1, SNP2, SNP3의 주효과들에 대한 도체중의 평균과 표준편차와 $P$-value

| $Marker$ | $Genotype$ | $numberofAnimals$ | $\bar{x} \pm s$ |
|---|---|---|---|
| SNP1 | AA | 16 | 294.19±26.999 |
| | AG | 75 | 303.97±31.227 |
| | GG | 138 | 313.61±34.369 |
| | $significance$(P$-value$) | | 0.020 |
| SNP2 | CC | 45 | 305.89±29.071 |
| | CT | 112 | 307.70±31.698 |
| | TT | 72 | 313.28±38.015 |
| | $significance$(P$-value$) | | 0.662 |
| SNP3 | AA | 54 | 307.74±35.971 |
| | AT | 118 | 313.17±33.754 |
| | TT | 57 | 301.95±28.786 |
| | $significance$(P$-value$) | | 0.703 |

$$Y_{ijkl} = \underset{(P)}{\mu} + \underset{(E)}{C_i + S_j + \beta age} + \underset{(G)}{M_k} + \varepsilon_{ijkl}, \tag{3.2}$$
$$i = 1, \ldots, c, \quad j = 1, \ldots, s, \quad k = 1, \ldots, m, \quad l = 1, \ldots, n.$$

여기서 $Y_{ijkl}$는 도체중이고 $C_i$(contemporary)는 도축계절과 장소를 같이 고려한 그룹으로 고정 효과이며 $S_j$는 sire 그룹의 랜덤 효과, $M_k$는 SNP 마커의 고정효과, $\beta$는 나이에 대한 회귀계수, $\varepsilon_{ijkl}$는 $N(0, \sigma^2)$인 확률변수이다. 그러나 우리가 관심을 가지는 부분은 도체중에 영향을 주는 것으로 환경효과보다 유전자효과에 관심이 있다. 따라서 유전자효과인 SNP가 도체중에 어떤 영향을 주는지에 대한 ANOVA 통계적 모형은 다음과 같다.

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \varepsilon_{ijkl}, \tag{3.3}$$
$$i = j = k = 1, 2, 3, \quad l = 1, 2, \ldots, n.$$

여기서 $Y_{ijkl}$는 도체중이고 $\alpha_i$는 SNP1, $\beta_j$는 SNP2, $\gamma_k$는 SNP3이 된다. 그리고 $\varepsilon_{ijkl}$는 $N(0, \sigma^2)$인 확률변수이다. 도체중에 대한 SNP의 영향을 보기 위하여 위의 ANOVA모형에 적용한 결과는 아래와 같다.

위의 표 3.1은 국가 후대 검정우 229두를 이용하여 SNP들의 주효과에 대한 도체중의 평균과 표준편차를 구하고 거기에 따라 ANOVA모형에서 SNP의 genotype간에 도체중의 평균에 차이가 있는지에 대한 $P$-value를 구한 것이고 표 3.2는 SNP들의 2개의 상호작용에 대하여 표 3.1과 같이 도체중의 평균과 표준편차를 구하고 거기에 따라 ANOVA모형에서 SNP의 genotype간에 도체중의 평균에 차이가 있는지에 대한 $P$-value를 구한 것이다. 결과를 보면 2개의 상호작용에 대한 효과는 모두 통계적으로 도체중에 유의한 차이가 나타나지 않았으며 1개의 주효과에 대한 것은 SNP1이 통계적으로 도체중에 유의한 차이(significance = 0.020)가 있다고 할 수 있다. genotype이 GG 타입일 경우 다른 타입보다

표 3.2: SNP1, SNP2, SNP3의 2개의 상호효과들에 대한 도체중의 평균과 표준편차와 $P$-value

| $Marker$ | $Genotype$ | $numberofAnimals$ | $\bar{x} \pm s$ |
|---|---|---|---|
| (SNP1*SNP2) | (AA , CC) | 5 | 307.00±25.436 |
| | (AA , CT) | 6 | 283.50±27.472 |
| | (AA , TT) | 5 | 294.20±27.689 |
| | (AG , CC) | 15 | 303.07±32.101 |
| | (AG , CT) | 35 | 304.03±30.014 |
| | (AG , TT) | 25 | 304.44±33.594 |
| | (GG , CC) | 25 | 307.36±28.860 |
| | (GG , CT) | 71 | 311.55±32.090 |
| | (GG , TT) | 42 | 320.81±40.211 |
| | $significance$(P$-value$) | | 0.322 |
| (SNP1*SNP3) | (AA , AA) | 3 | 292.33±11.015 |
| | (AA , AT) | 6 | 297.83±39.163 |
| | (AA , TT) | 7 | 291.86±21.965 |
| | (AG , AA) | 19 | 296.84±34.012 |
| | (AG , AT) | 38 | 312.74±29.431 |
| | (AG , TT) | 18 | 293.00±27.903 |
| | (GG , AA) | 32 | 315.66±37.044 |
| | (GG , AT) | 74 | 314.64±35.502 |
| | (GG , TT) | 32 | 309.19±29.205 |
| | $significance$(P$-value$) | | 0.243 |
| (SNP2*SNP3) | (CC , AA) | 10 | 311.10±28.521 |
| | (CC , AT) | 26 | 308.73±32.189 |
| | (CC , TT) | 9 | 291.89±14.641 |
| | (CT , AA) | 31 | 309.39±35.861 |
| | (CT , AT) | 50 | 306.78±32.783 |
| | (CT , TT) | 31 | 307.48±25.918 |
| | (TT , AA) | 13 | 301.23±42.748 |
| | (TT , AT) | 42 | 323.52±34.103 |
| | (TT , TT) | 17 | 297.18±37.323 |
| | $significance$(P$-value$) | | 0.135 |

더 도체중이 큰 것을 알 수 있다. 하지만 우리는 앞에서 경제형질인 도체중에 영향을 주는 유전자가 개별 SNP유전자가 아닌 여러개의 복잡한 SNP유전자의 상호작용이 더 영향을 준다고 믿고 있었다. 그러나 ANOVA를 이용한 방법에서는 도체중에 상호작용이 주는 영향은 없는 것으로 나타났다. ANOVA는 도체중에 영향을 주는 요인을 하나의 요인과 상호작용요인을 한꺼번에 같이 적용하는 모형을 사용하므로 하나의 효과이면 하나의 효과만, 상호작용이면 상호작용효과만 고려하는 방법인 MDR 모형에 적용하여 보겠다.

### 3.3. Multifactor Dimensionality Reduction(MDR) 방법을 이용한 분석

앞에서 말한 것과 같이 여러 개의 유전자의 상호작용이 경제형질에 더 영향을 준다고 할 수 있다. 따라서 여러 상호작용에 대한 high-order 차수의 데이터로 복잡한 관계를 밝힐 수 있는 MDR 방법을 이용하여 분석을 해보자. MDR 방법은 case와 control data에서 여러 유전자의 상호작용을 찾을 수 있다. 도체중에 영향을 주는 SNP들의 상호작용을 알아보기 위한 방법으로 MDR 방법을 이용한다. MDR 방법은 case와 control 두 부분으로 나누어져야 하고 case와 control이 1:1로 같아야한다. 그러나 원 데이터에서는 도체중 변수가 연속인 데이터이다. 그래서 그대로 MDR 방법에 적용할 수 없다. 따라서 첫 단계에서는 도체중 변수를 데이터마이닝기법 중 CART 방법으로 case-control로 나눈 뒤 Ritchie 등 (2003)과 Hahn 등 (2003)의 MDR 방법에 적용한다.

**Step 1.** 도체중 변수를 CART 방법으로 이분화를 시킨다. 그림 3.1의 STEP 1을 보면 CART 방법에서 도체중 변수를 기준으로 등급이 case(high)와 control(low)로 나 누어진다. 여기서 case그룹이 54개 control그룹이 175개로 나누어지는데 이렇게 나누어진 데이터는 1:1이 되지 않으므로 SPSS clementin10.1을 이용하여 case 그룹 175개를 기준으로 균형 증폭 시킨다. 이렇게 증폭되어진 350두를 이용하여 다음 Step들에 적용한다.

**Step 2.** 데이터를 랜덤으로 10개의 같은 크기로 나눈다. 그리고 그 중 9개를 training set으로 정하고 1개를 testing set으로 정한다.

**Step 3.** 3개의 SNP들로부터 2개의 SNP조합 중 하나를 선택한다.

**Step 4.** 선택된 SNP조합에서 SNP의 각각 수준을 기초로 한 개체들을 multifactor classes 또는 cells에 기술한다. 위의 그림의 STEP 4와 같이 2개의 조합이므로 $3^2 = 9$개의 셀을 가진다. 각각 9개의 셀에 case의 도수와 control의 도수를 적는다.

**Step 5.** Case와 control의 비를 구하여 1보다 크거나 같으면 high-risk, 1보다 작으면 low-risk라 한다. 위의 그림에서처럼 1행 1열의 경우 case와 control의 비가 1.333이므로 high-risk가 된다.

**Step 6.** 2개의 SNP의 조합 전부에서 데이터의 9/10인 traing set에서 잘못 분류된 비율인 misclassification error(ME)를 구한다. 위의 그림에서처럼 ME의 값이 가장 작은 SNP1과 SNP2의 조합을 선택한다.

**Step 7.** Training set에서 high-low로 나눈 표를 나머지 1/10의 데이터인 testing set을 이용하여 잘못 분류된 비율인 prediction error(PE)를 구한다.

이처럼 2개의 SNP의 상호작용에 대하여 Step 2~Step 7의 과정을 10번 반복해서 나온 ME와 PE의 평균과 10번의 반복과정에서 나온 값을 기준으로 CVC를 구한 결과와 같은 방법으로 하나의 SNP에 대한 ME와 PE의 평균과 CVC의 결과는 표 3.3과 표 3.4와 같이 나타났다.

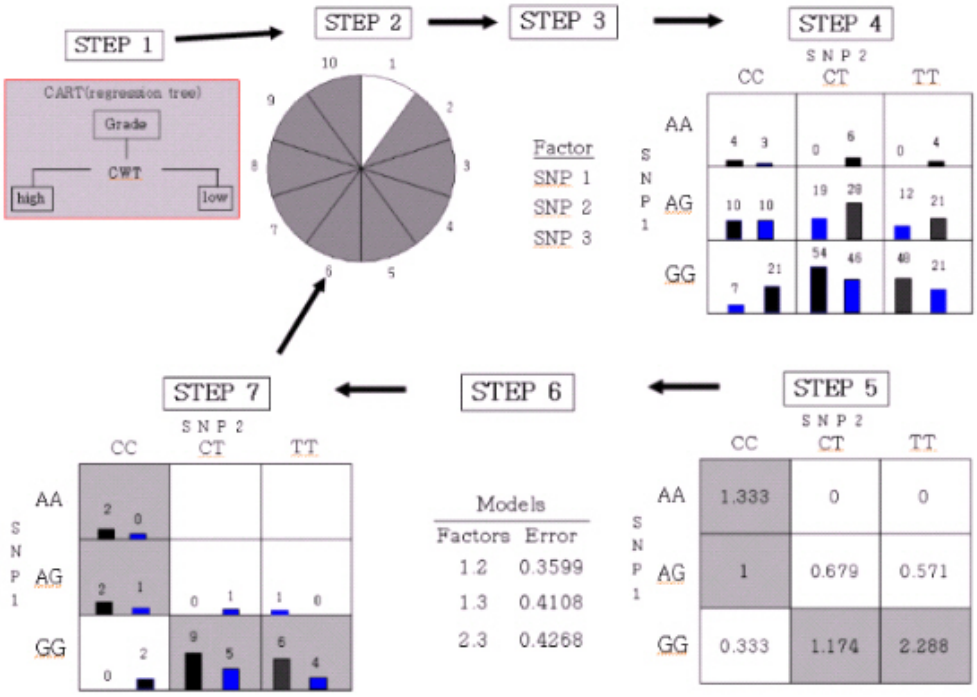


그림 3.1: 도체중에 대하여 SNP들을 이용한 MDR 방법의 적용과정

표 3.3: 도체중에 대한 SNP 각각 조합의 average ME와 average PE 결과

| $Number of factors$ | $Marker$ | $averageME$ | $averageME$ |
|---|---|---|---|
| 1 | SNP1 | 0.4299 | 0.4412 |
| | SNP2 | 0.4478 | 0.4971 |
| | SNP3 | 0.4322 | 0.4588 |
| 2 | SNP1*SNP2 | 0.3713 | 0.3853 |
| | SNP1*SNP3 | 0.4131 | 0.4353 |
| | SNP2*SNP3 | 0.4226 | 0.4706 |

위의 표 3.3은 MDR과정을 10번 반복한 후 각각의 SNP 조합에 대한 ME와 PE의 평균을 구한 값이다. Best model은 average ME와 average PE가 가장 낮은 것을 선택하는데 위의 표를 보면 하나의 요인에 대한 효과에서 SNP1의 average ME가 0.4299. PE가 0.4412로 가장 낮게 나타났고 두개의 요인에 대한 효과에서는 SNP1*SNP2의 average ME가 0.3713, PE가 0.3853으로 가장 낮은 걸로 나타났다. 따라서 하나의 요인에서는 SNP1이 선택되었고 두 개의 요인에서는 SNP1*SNP2가 선택되었다. 그리고 SNP1과 SNP1*SNP2를 보면 상호작용 효과인 SNP1*SNP2의 average ME와 PE가 더 낮은 것을 알 수 있다. 따라서 하나의 요인에 대한 효과보다는 상호작용에 대한 효과가 더 많다는 것을 알 수 있다. 다음으로

표 3.4: 도체중에 대한 하나의 유전자효과와 두 개의 상호작용 효과의 CVC 결과

| Number of factors | 1 | | | 2 | | |
|---|---|---|---|---|---|---|
| Cross Validation Interval | SNP1 ME | SNP2 ME | SNP3 ME | SNP1*SNP2 ME | SNP1*SNP3 ME | SNP2*SNP3 ME |
| #1 of 10 | 0.4299 | 0.4522 | 0.4395 | 0.3599 | 0.4108 | 0.4268 |
| #2 of 10 | 0.4236 | 0.4459 | 0.4299 | 0.3758 | 0.4204 | 0.4204 |
| #3 of 10 | 0.4299 | 0.4427 | 0.4331 | 0.3758 | 0.4076 | 0.4045 |
| #4 of 10 | 0.4236 | 0.4395 | 0.4299 | 0.3694 | 0.4108 | 0.4236 |
| #5 of 10 | 0.4268 | 0.4459 | 0.4140 | 0.3599 | 0.3949 | 0.4045 |
| #6 of 10 | 0.4363 | 0.4490 | 0.4236 | 0.3790 | 0.4236 | 0.4236 |
| #7 of 10 | 0.4427 | 0.4490 | 0.4459 | 0.3694 | 0.4172 | 0.4363 |
| #8 of 10 | 0.4363 | 0.4682 | 0.4522 | 0.3726 | 0.4076 | 0.4395 |
| #9 of 10 | 0.4299 | 0.4395 | 0.4268 | 0.3758 | 0.4172 | 0.4236 |
| #10 of 10 | 0.4204 | 0.4459 | 0.4268 | 0.3758 | 0.4204 | 0.4236 |
| count | 7 | 0 | 3 | 10 | 0 | 0 |

CVC결과 (표 3.4)를 살펴보자.

이는 10번의 cross-validation을 시행할 때 각 시행에서 선택된 best model을 카운트한 값이 CVC라고 한다. 위의 표 3.4를 보면 #1 of 10의 시행에서 ME가 SNP1이 0.4299, SNP2가 0.4522, SNP3가 0.4395으로 SNP1이 가장 작으므로 SNP1에 대하여 count +1을 한다. 총 10번을 시행 했을 때 표 3.4와 같이 SNP1이 각 시행에서 가장 좋은 모형으로 선택된 것이 7번이기 때문에 SNP1의 CVC는 7이 된다. SNP1*SNP2 역시 각 시행에서 가장 좋은 모형으로 10번이 선택되었기 때문에 CVC는 10이 된다. 따라서 표 3.3과 표 3.4를 보면 최종 best model은 요인이 한 개 일 때는 average ME와 average PE가 가장 낮고 CVC가 가장 높은 SNP1 set을 선택되고 요인이 두 개 일 때는 average ME와 average PE가 가장 낮고 CVC가 가장 높은 SNP1*SNP2 set이 최종 best set으로 선택되었다.

## 4. 결론 및 토의

우리는 복합질병에 대한 위험 또는 가축의 경제특성의 여러 유전자에 관련된 polymorphism의 조합을 이용하여 모수적인 방법인 ANOVA와 비모수적인 방법인 MDR을 소개하였다. 이들 방법을 적용하여 도체중에 영향을 주는 SNP들을 규명하려고 시도한 결과 모수적인 방법인 ANOVA에서는 SNP1이 도체중에 유의한 차이가 있었고 상호작용에 대한 영향은 없는 것으로 나타났다. 하지만 비모수적인 방법인 MDR방법에서는 하나의 SNP에서는 SNP1이 도체중에 영향을 많이 주었으며 두 개의 상호작용에서는 SNP1*SNP2가 영향을 많이 주는 것을 알 수 있었다. 또한 SNP1과 SNP1*SNP2의 ME와 PE를 비교한 결과 SNP1인 하나의 유전자의 효과 보다는 SNP1*SNP2와 같이 상호작용에 의한 유전효과가 도체중 경제형질에 더 많은 영향을 주는 인자로 밝혀졌다. 아울러 CVC 검정결과에서도 SNP1*SNP2 인자가 최종모형임이 밝혀졌다. 따라서 한우의 도체중에 영향을 주는 SNP를 규명하는 문제에서 모수적인 방법인 ANOVA모형에서 나타나지 않은 주요한 상호작용에 관계된 인자

가 MDR 방법에 의해 규명되어졌다.

한편 MDR 방법에서 문제점은 데이터가 case-control로 이분형이 되어야하기 때문에 연속형 자료는 사용할 수 없다는 것이다. 이런 문제점을 해결하기 위해서 본 연구에서는 도체중 변수와 같은 연속형 자료는 앞 절에서와 같이 도체중 변수를 CART 기법으로 미리 case와 control로 나눈 뒤에 MDR에 적용할 수 있었다. 그리고 또 다른 문제점으로 MDR방법은 상대적으로 적은 표본 수를 가진 경우에 상호작용을 규명하는 비모수적인 방법이기 때문에 유전자의 수가 많을 경우 각 셀에 관측값이 없는 빈 셀이 생길 수 있다. 예를 들면 SNP 유전자의 수가 10개 이상일 경우 SNP 유전자 10개 상호작용에서의 효과를 보면 인 59049개의 셀로 나누어지기 때문에 많은 셀 안에 관측값이 없는 경우가 생긴다. 본 연구에서는 SNP 유전자의 수가 3개이기 때문에 MDR 방법에 적용할 수 있었으나, 많은 SNP 유전자들의 경우 상호작용규명에 문제가 있음은 더 연구되어야 할 부분이다.

## 참고문헌


Barendse, W., Bunch, R., Thomas, M., Armitage, S., Baud, S. and Donaldson, N. (2004). The TG5 thyroglobulin gene test for a marbling quantitative trait loci evaluated in feedlot cattle, *Australian Journal of Experimental Agriculture*, **44**, 669–674.

Bastione, 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 Geredity*, **58**, 82–92.

Chung, Y. J., Lee, S. Y., Park, T. S. (2005). Multifactor dimensionality reduction in the presence of missing observations, In *Proceedings of the Autumn Conference, The Korea Statistical Society,* 31–36.

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*, 2nd ed., 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. (2007). Identification of candidate SNP for economic traits on chromosome 6 in Korean cattle, *Animal Genetics*, submitted.

Page, B. T., Casas, E., Quaas, R. L., Thallman, R. M., Wheeler, T. L., Shackelford, S. D., Koohmaraie, M., White, S. N., Bennett, G. L., Keele, J. W., Dikeman, M. E. and Smith, T. P. L. (2004). Association of markers in the bovine CAPNI gene with meat tenderness in large crossbred populations that sample influential industry sires, *Journal of Animal Science*, **82**, 3474–3481.

Ritchie, M. D., Hahn, L. W. and Moore, J. H. (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity, *Genetic Epidemiology*, **24**, 150–157.

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. P. L. (2005). Linkage mapping bovine EST-based SNP, *BMC Genomics*, **6**,74–84.

[ 2007년 6월 접수, 2007년 11월 채택 ]

# Main SNP Identification of Hanwoo Carcass Weight with Multifactor Dimensionality Reduction(MDR) Method*

Jea-Young Lee[1] Dong Chul Kim[2]

Abstract

It is commonly believed that disease of human or economic traits of livestock are caused not by single gene acting alone, but by multiple genes interacting with one another. This issue is difficult due to the limitations of parametric statistical method like as logistic regression for detection of gene effects that are dependent solely on interactions with other genes and with environmental exposures. Multifactor dimensionality reduction (MDR) nonparametric statistical method, to improve the identification of single nucleotide polymorphism (SNP) associated with the Hanwoo(Korean cattle) carcass cold weight, is applied and compared with ANOVA results.

*Keywords:* Single nucleotide polymorphism(SNP), multifactor dimensionality reduction(MDR), carcass cold weight.

* This research was supported by Gyeongbuk Hanwoo cluster(Ministry of Agriculture and Forestry) Republic of Korea.

1) Corresponding author. Professor, Dept. of Statistics, Yeungnam University, Kyungsan 712-749, Korea. E-mail: jlee@yu.ac.kr

2) Graduate, Dept. of Statistics, Yeungnam University, Kyungsan 712-749, Korea. E-mail: kitty2142@ynu.ac.kr