응용통계연구 제 15권 1호, 2002년, pp. 85-105

유전자검사자료의 통계분석을 위한 수량화 및 그래프 방법 *

박미라[1)]

요 약

본 연구에서는 유전자 검사자료에서 각 유전자좌내 및 유전자좌간의 대립형질들간의 관계를 파악하기 위한 탐색적 방법을 고려하였다. 이를 위해 유전자데이터를 재배열한 후 대응분석 및 다중대응분석의 알고리즘을 적용하여 이를 수량화, 그래프화하는 방법 및 대응행렬도를 적용한 방법을 제안하였다. 이러한 수량화 및 그래프 결과가 하디-와인버그 평형검정 및 연관균형검정 결과와 어떤 관계가 있는지 알아보고 실제 한국인 집단에 대한 STR 유전자좌 자료를 이용하여 결과를 비교하였다.

주요용어: 하디-와인버그 평형, 연관균형, 대응분석, 다중대응분석, 대응행렬도, 다중대응행렬도.

1. 서 론

유전자검사를 이용한 개인식별(individual identification)기술은 PCR (polymerase chain reaction)기술과 STR 유전자좌(short tandem repeats loci)의 자동 대립형질분석방법의 개발로 보다 신속하고 비용이 적게 들면서도 적은 표본으로 높은 정밀도를 얻을 수 있도록 급진전되었다(cf. Alford et al., 1994). 개인식별을 위해서는 분석할 유전자를 선택하여 검사체계를 구성하고 미리 조사된 대립유전자 빈도를 이용하여 유전자 증거를 통계적으로 정량화하고 증거력을 평가하는 절차를 거치게 된다. 이 때 특정 분석집단에서 발현되는 유전적 관련성에 따라 결과가 영향을 받게 되므로 분석하고자 하는 집단의 유전적 특성에 관한 연구가 선행되어야 한다. 적절한 검사체계의 구성을 위해서 필요한 사항 중 하나는 각 유전자좌(locus)의 대립형질 빈도(allele frequency)를 조사하는 것인데 이 때 유전자좌내에서의 대립형질들 및 여러 유전자좌들에서의 대립형질 사이의 관계를 조사하는 것은 매우 중요하다. 일반적으로 이 문제는 대립형질쌍의 빈도가 각각의 빈도의 곱으로 표현될 수 있느냐 하는 독립성에 관한 것으로 이것이 보장되지 않는다면 각 경우마다 유전자증거의 평가를 위한 확률계산이 달라져야 하므로 사전조사를 통해 검증할 필요가 있다. 하나의 유전자좌에서 두 대립형질간의 문제는 하디-와인버그 평형(Hardy-Weinberg equilibrium)을 유지하느냐에 대한 문제가 되며, 두 개의 유전자좌에서 나온 대립형질간의 관계는 연관균형(linkage equilibrium)여부에 관한 문제로 귀착된다. 최근에도 각 민족들에 대해 유전좌별로 이러한 문제를 다룬 연구결과가 계속 진행중이며, 각 민족간의 비교연구도 수행되고 있다(Budowle

* 이 논문은 1999년도 한국학술진흥재단의 연구비에 의하여 연구되었음(KRF-1999-003-D00072)

1) (301-832) 대전시 중구 용두2동 143-5, 을지의과대학교 의예과, 조교수
E-mail: mira@emc.eulji.ac.kr

et al., 1997;Gehrig et al., 1999;한길로 외, 1999; Fregeau et al., 1998; 이재원·박미라, 2000, Lee et al., 2001).

이러한 연구들은 모두 최종적인 검정의 결과(p-값)만을 제공하고 있다. 그러나 각 대립형질들간의 관계를 2차원 그래프로 표현할 수 있다면 연구자가 집단의 특성에 대한 전반적인 경향을 파악하는데 큰 도움을 줄 수 있을 것이다. 탐색적인 다변량분석방법으로서 행렬도(Gabriel,1971)나 일본의 수량화방법(駒澤 勉,1982), 대응분석(Greenacre, 1984)등을 생각할 수 있다. 여기서는 유전적 특성들을 수량화하고 그들간의 잠재적인 구조를 탐색하기 위한 방법으로 유전자데이터를 재배열한 후 대응분석의 알고리즘을 적용하여 2차원 그래프로 표시하는 방법을 제안하였다. 또한 이를 행렬도식으로 표현하는 방법도 제시하였다. 이 결과로서 유전자좌내, 유전자좌간의 대립형질간의 관계 및 대립형질의 상대적인 중요도 등을 파악할 수 있다. 2절에서는 유전학에서의 평형개념을 설명하고 3절에서는 유전자분석자료의 수량화 및 그래프화를 위한 방법을 설명하겠다. 4절에서는 한국인의 STR 자료를 이용한 분석결과를 보이겠다.

2. 유전학에서 평형의 의미

집단유전학에서 가장 중요한 이론 중 하나는 하디-와인버그법칙이다. 이 법칙은 자연선택(selection), 돌연변이(mutation), 이주(migration)가 없이 임의교배가 이루어지는 집단에서 특정 유전자좌에 대한 대립형질(allele) 또는 유전자형(genotype)의 빈도가 세대에 관계없이 늘 일정한 값을 유지하며, 대립유전자빈도와 유전자형빈도간에는 단순한 관계가 있다는 것이다(Hardy, 1908;Crow,1988). r개의 대립유전자 $A_1, A_2, \cdots A_r$을 가진 유전자좌가 있고 A_i 의 대립유전자비율을 f_i라고 하자. 하디-와인버그 평형을 유지한다면, 동형접합자 A_iA_i의 유전자비율은 f_i^2 이 되고 이형접합자 A_iA_j의 비율은 $2f_if_j$가 된다. 그러나 이 이론은 이상적이지만 대부분의 자연적인 집단에서는 배우자의 선택에서도 특정한 경향이 있게 마련이며, 구성원의 이주나 돌연변이 등이 발생하여 집단의 대립유전자는 시간이 지남에 따라 어느 정도는 변하게 된다. 이러한 요인들로 인하여 집단 내에 대립유전자비율을 달리하는 아집단(subpopulation)이 형성된 집단을 이형집단(heterogeneous population)이라고 하며 이러한 현상을 월룬트효과(Wahlund effect)라고 한다. 유전자 감식에 있어서 어떤 유전자좌가 하디-와인버그평형을 충족하지 않는다면 각 경우마다 유전자증거의 평가를 위한 확률계산이 달라져야 하므로 이를 파악하는 것은 매우 중요하다.

이러한 평형을 만족하는지에 대한 여러 검정방법들이 개발되어 왔다(Ward and Sing, 1970; Emigh, 1980; Guo and Thomson,1992; Shoemaker et al., 1998). 흔히 쓰이는 방법으로 평형상태에서의 유전자형의 기대빈도수를 계산하는 적합도검정이 있다.대립유전자 i 에 대한 추정비율 $\hat{f}_i$는

$$\hat{f}_i = f_{ii} + \frac{1}{2}\sum_{j \neq i} f_{ij}$$

이 된다. 여기서 f_{ii}는 대립유전자 i 두 개로 이루어진 동형접합자 유전자형의 관찰비율을, f_{ij}는 대립유전자 i 와 j를 하나씩 포함하는 이형접합자 유전자형의 관찰비율을 나타낸다.

이 때 다음과 같은 통계량

$$\chi^2 = \sum_i \frac{(x_{ii}^* - n\hat{f}_i^2)^2}{n\hat{f}_i^2} + \sum_i \sum_{i \neq j} \frac{(x_{ij}^* - 2n\hat{f}_i\hat{f}_j)^2}{2n\hat{f}_i\hat{f}_j} \tag{2.1}$$

이 자유도 $r(r-1)$인 카이제곱분포를 따른다는 것을 이용하여 검정하게 된다. 여기서 x_{ii}^*는 대립유전자 i 두 개로 이루어진 동형접합자 유전자형의 관찰빈도를, x_{ij}^*는 대립유전자 i와 j를 하나씩 포함하는 이형접합자 유전자형 관찰빈도를 나타낸다. 또한 r은 대립유전자수이며 $n = \sum_i \sum_{i \neq j} x_{ij}^*$이다. 표본수가 적은 경우 기대값이 작은 셀들이 많이 생길 때에는 유전자형에 대해 모든 가능한 결과의 확률을 구하여 검정하는 정확검정(exact test)방법들을 사용할 수 있다(Guo and Thompson, 1992). 여기서도 이 두 가지 방법에 의한 p-값을 계산하여 그래프와 함께 제시하게 될 것이다.

이와 마찬가지로 어떤 한 유전자좌에 있는 대립유전자의 발생빈도가 다른 유전자좌에 있는 대립유전자의 발생빈도에 영향을 주는지 검토가 필요한데 이는 집단이 연관 불균형 상태를 유지하고 있는가에 대한 검정이 된다. 연관 불균형은 두 유전자좌가 동일 염색체상에 존재할 경우는 물론 서로 다른 염색체상에 존재할 경우에도 있을 수 있다. 특히 집단의 규모가 작거나 무작위적인 결혼(random mating)이 이루어지지 않는 집단에서 발생할 확률이 높다. 세대가 거듭될수록 연관 불균형 상태는 연관 균형 상태로 진행을 하게 되는데 서로 다른 염색체에 존재하는 유전자좌들이나 재조합 확률이 높은 유전자좌들은 상대적으로 빨리 연관균형 상태에 이르게 된다. p_{AB}가 두 유전자좌 A, B에서 동시에 나타나는 대립유전자쌍의 관측확률이고 p_A와 p_B가 두 유전자좌 각각의 대립유전자비율이라 하면 연관불균형 계수

$$D_{AB} = p_{AB} - p_A p_B$$

로 정의된다. 여기서 연관 균형을 이룬 상태라면 D_{AB}는 0 이 될 것이므로 연관불균형에 대한 검정은 유전자좌간 대립유전자의 관측값과 기대값 차이를 0으로 할 수 있는가에 대한 검정이다. 역시 적합도 검정과 정확검정법 등이 가능하다(Zaykin et al., 1995).

3. 대립형질의 수량화 및 그래프 방법

3.1. 하나의 유전자좌에서의 대립형질의 수량화와 그래프 방법

DNA검사 결과 각 조사대상자는 하나의 유전자좌에 대해 각각 2개의 대립형질(allele)을 보이게 된다. 데이터는 그림 3.1과 같은 형식으로 얻어지며 여기서 x_{ii}^*는 대립유전자 i 두 개로 이루어진 동형접합자 유전자형의 관찰빈도를, x_{ij}^* 는 대립유전자 i 와 j를 하나씩 포함하는 이형접합자 유전자형 관찰빈도를 나타낸다($j > i$). r개의 대립형질이 있을 때 $x_i^* = x_{ii}^* + x_{i+}{}^* = x_{ii}^* + \frac{1}{2}\sum_{j=1}^{r} x_{ij}^*$ 로 하면 x_i^*는 표본에서 i번째 대립유전자의 수가 된다.

이 데이터를 $x_{ii} = x_{ii}^*, x_{ij} = x_{ji} = x_{ij}^*/2$ 로 하여 $r \times r$의 대칭행렬 $X = (x_{ij})$로 재표현하자. 이제 각 대립형질의 수량화는 행렬 X의 행(열)의 수량화문제로 생각할 수 있다. 여기서는 이를 위한 방법으로 대응분석(correspondence analysis)의 알고리즘을 쓰기로 한다.

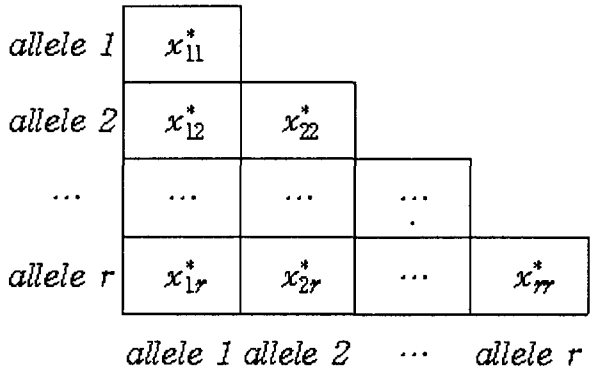

그림 3.1: r개의 대립형질이 있을 때의 데이터 형태

일반적으로 $r \times c$크기의 자료행렬 $X = (x_{ij})$를 원소들의 전체합으로 나눈 원소들로 이루어지는 대응행렬을 $F = (f_{ij})$라고 할 때 대응분석은 이들 r차원과 c차원의 가중유클리드 공간에서 정의되는 좌표점들을 카이제곱거리의 성질을 가지면서 차원축소된 그래프로써 데이터를 표현하는 주성분분석의 일종이라고 할 수 있다. 이의 수량화 절차는 다음과 같은 행렬 G의 비정칙치분해(singular value decomposition)

$$G = D_r{}^{-1/2}(F - rc')D_c{}^{-1/2} = UD_\lambda V' \tag{3.1}$$

로부터 구해질 수 있다. 여기서 $r = (f_{1+}, \cdots, f_{r+})'$, $c = (f_{+1}, \cdots, f_{+c})'$, $f_{i+} = \sum_j f_{ij}$, $f_{+j} = \sum_i f_{ij}$이고 D_r과 D_c는 r과 c를 대각원소로 하는 대각행렬이다. 또한 D_λ는 G의 비정칙치를 대각원소로 하는 대각행렬이며 $U'U = V'V = I$이다. c차원 공간상의 r개의 프로파일과 r차원 공간상의 c개 프로파일을 저차원그래프에 나타내기 위한 수량화값으로 통상적인 대응분석에서 이용되는 것은

$$A = D_r{}^{-1/2}UD_\lambda, \quad B = D_c{}^{-1/2}VD_\lambda$$

으로 이들의 처음 두 열을 좌표점으로 하여 2차원 그래프를 얻게 된다. 또는 이의 축의 척도를 조금씩 달리하여 쓰기도 한다(최용석, 1993; 허명회, 1998). 행좌표점간(또는 열좌표점간)의 유클리드 거리는 행프로파일간(또는 열프로파일간)의 카이제곱거리를 의미한다. 또한 행과 열간의 독립성 검정을 위한 카이제곱검정통계량은

$$\chi^2 = \sum_i \sum_j \frac{(x_{ij} - x_{i+}x_{+j}/x_{++})^2}{x_{i+}x_{+j}/x_{++}} = n \sum_i f_{i+} \ (r_i - c)' D_c{}^{-1}(r_i - c)$$

이다. 여기서 $x_{i+} = \sum_j x_{ij}$, $x_{+j} = \sum_i x_{ij}$, $x_{++} = \sum_i \sum_j x_{ij}$, $r_i = (f_{i1}/f_{i+}, \cdots, f_{ic}/f_{i+})'$로 행프로파일이다. 따라서 기하적으로 볼 때 χ^2/n(총 inertia)은 행프로파일들과 행중심점(row centroid) c간의 제곱카이제곱거리(squared chi-squared distance)의 가중평균이 되며 이는 동질성 검정의 결과와도 같게 된다. 이 때 행좌표점과 열좌표점사이의 거리에는 아무런 기

하적 의미가 없고 다만 두 좌표점이 같은 거리에 위치할 때 이들 행범주와 열범주간에 대응관계가 있다고 해석하게 된다(Greenacre and Hastie, 1987).

유전자 자료에서는 대응행렬이 대칭이므로 행의 수량화 또는 열의 수량화 하나만을 얻는 것으로 충분하며 행렬 G가 양정칙(positive definite)행렬일 경우에는 두 해가 동일하게 된다. 또한 여기서 중심화된 대응행렬을 이용하였으므로 χ^2/n는 좌표점들의 원점으로부터의 거리제곱의 가중합이 된다. 이 때 i번째 대립형질에 대한 가중치는 유전자비율 f_{i+}이 된다. 따라서 그래프에서 유사한 위치에 찍히는 대립형질은 유사한 행(열)프로파일을 가지는 것을 의미하고, 카이제곱통계량이 유의하면(즉, 하디-와인버그 평형이 깨진다면) 좌표점들이 중심으로부터 멀리 떨어져서 찍히는 경향이 있다.

그런데 식 (3.1)로부터 행렬 G의 (i,j)번째 원소는

$$g_{ij} = \frac{(x_{ij} - \frac{x_{i+}x_{+j}}{x_{++}})}{\sqrt{x_{i+}x_{+j}}}$$

가 되어 독립성검정을 위한 카이제곱통계량의 (i,j)번째 칸의 기여도 $\times n$이 된다. 따라서 통상적인 좌표점 대신

$$A^* = UD_\lambda^\alpha, \quad B^* = VD_\lambda^{1-\alpha}$$

를 사용하면 $G = A^*B^{*'}$이 된다(Gower and Hand, 1996). 즉, a_i^* 를 A^*의 i번째 행, b_j^*를 B^*의 j번째 행이라고 할 때

$$g_{ij} = a_i^{*'} b_j^*$$

이 되어 이들의 내적이 카이제곱통계량의 기여도에 비례하는 성질을 갖게 된다. 여기서는 편의상 이를 "대응행렬도"라고 부르겠다. 유전자자료에서 행렬 G가 양정칙행렬일 때에는 행좌표점과 열좌표점이 같아지므로 $\alpha = 1/2$로 하여 $UD_\lambda{}^{1/2}$의 처음 두 열을 수량화값으로 그래프를 그리면 이들간의 내적이 모두 의미를 갖게 되어 내적이 큰 대립형질들이 하디-와인버그 평형을 깨는데 역할을 많이 하는 것으로 해석할 수 있다. 또한 두 점이 이루는 각이 예각일 때는 카이제곱검정통계량이 양수(즉, 관측치보다 기대치가 큼)라는 것을 의미하고 둔각일 때에는 부호가 반대가 됨을 의미한다. 직각을 이루는 대립형질들은 서로 독립임을 나타낸다. 단 이때에는 통상적인 대응분석과는 달리 점들간의 거리는 아무런 의미가 없게 된다. 행렬 G가 음의 고유치를 갖는 경우에는

$$G = UD_\lambda KU'$$

로 표현할 수 있다. 여기서 K는 대각행렬로서 해당하는 고유치가 양수일 때 $k_{ii} = 1$, 음수일 때 $k_{ii} = -1$로 정의된다. 따라서 이 경우에는 하나의 그래프가 아니고 행$(A^* = UD_\lambda{}^{1/2})$과 열$(B^* = UKD_\lambda{}^{1/2})$의 그래프 모두를 그리고 행점과 열점간의 내적을 해석하여야 한다.

3.2. 두 유전자좌에서의 대립형질의 수량화와 그래프 방법

이제 두 개의 유전자좌 X, Y 에서 각 대립형질들의 관계를 수량화하는 방법을 생각하여보자. 유전자좌 X가 k_1개, 유전자좌 Y가 k_2개의 대립형질을 가진다고 하면 자료를 다음

과 같은 다중표시행렬(multiple indicator matrix)로 표현할 수 있다.

$$Z = (Z_{11} : Z_{12} : Z_{21} : Z_{22})$$

$$n^* \times k \quad n^* \times k_1 \; n^* \times k_1 \; n^* \times k_2 \; n^* \times k_2$$

여기서 Z_{11}, Z_{12}는 유전자좌 X에 해당하는 표시행렬이고, Z_{21}, Z_{22}는 유전자좌 Y에 해당하는 표시행렬이다. 예를 들어 각 유전자좌가 두 개씩의 대립형질을 가지고 있고 어떤 사람이 유전자좌 X 에서 (1,1) Y에서 (2,2)를 가졌다면 그 사람의 데이터는

$$(1\,0 : 1\,0 : 0\,1 : 0\,1)$$

이 된다. 그런데 (1, 1)과 (1, 2)를 가진 사람은 (1 0 : 1 0 : 1 0 : 0 1) 또는 (1 0 : 1 0 : 0 1 : 1 0)로 표시할 수 있게 되고 (1,2)과 (1,2)를 가진 사람은 (1 0 : 0 1 : 1 0 : 0 1) (0 1 : 1 0 : 1 0 : 0 1) (1 0 : 0 1 : 0 1 : 1 0) (0 1 : 1 0 : 0 1 : 1 0)의 네가지 표현이 모두 가능하다. 여기서는 이러한 점을 고려하여 모든 가능한 경우에 대해 표기하고 관측돗수에 비례하도록 같은 대립형질을 가진 데이터는 반복하여 입력하기로 한다. 즉 이 예에서와 같이 (1,1,2,2), (1,1,1,2), (1,2,1,2)의 세 사람이 있다면 최종 데이터는

10100101
10100101
10100101
10100101
10101001
10101001
10100110
10100110
10011001
01101001
10010110
01100110

이 된다. 따라서 n명에 대한 자료가 있다면 $n^* = 4n$이다. 여기에 다중대응분석의 알고리즘을 이용하여 각 대립형질을 수량화하는 방법을 고려할 수 있다. 단순대응분석을 일반화하는 다중대응분석방법으로는 여러 가지 접근 방법을 생각할 수 있는데 Park and Huh(1996)의 일반화정준상관분석의 기하적 해석과 관련하여 허명회(1999)는 수량화를 위한 목표로 다음과 같은 정식화를 고려하였다.

$$min \sum_{i,j,l,m} || \, Z_{ij}c_{ij} - Z_{lm}c_{lm} \, ||^2 \, /n^* \tag{3.2}$$

이 때의 조건식은 $\sum_{ij} c_{ij} Z_{ij}{}' Z_{ij} c_{ij}/n^* = m$과 $1_{n^*}{}' Z_{ij}{}' c_{ij}/n^* = 0$ 이다$(\forall i, j)$. 즉, $Z_{11}c_{11}, \cdots, Z_{mm}c_{mm}$을 가능한 한 근접시키고자 하는 것을 수량화의 목표로 한 것이다.

이의 해는 다음과 같은 버트행렬(Burt matrix)로부터 생성될 수 있다.

$$\begin{pmatrix} Z_{11}'Z_{11} & Z_{11}'Z_{12} & Z_{11}'Z_{21} & Z_{11}'Z_{22} \\ Z_{12}'Z_{11} & Z_{12}'Z_{12} & Z_{12}'Z_{21} & Z_{12}'Z_{22} \\ Z_{21}'Z_{11} & Z_{21}'Z_{12} & Z_{21}'Z_{21} & Z_{21}'Z_{22} \\ Z_{22}'Z_{11} & Z_{22}'Z_{12} & Z_{22}'Z_{21} & Z_{22}'Z_{22} \end{pmatrix}$$

여기서 $Z'Z$는 대칭행렬로 대각블럭 $Z_{ij}'Z_{ij}$는 원소 Z_{ij}의 열합, 즉 대립형질의 관측수로 이루어진 대각행렬이고, 비대각블럭 $Z_{ii}Z_{ll}$에는 i번째 유전자좌와 l번째 유전자좌에 의한 2원 분할표가 들어가게 된다. 이의 수량화 절차는 다음과 같은 행렬 H의 고유치분해

$$H = {D_B}^{-1/2} B {D_B}^{-1/2} = U D_\lambda U' \tag{3.3}$$

에서 구할 수 있다. 여기서 $B = Z'Z/n^*$이며 D_B는 B의 대각원소로 이루어진 대각행렬이다. (3.2)의 기준에 따르면 수량화값은

$$A = \sqrt{m}\ {D_B}^{-1/2} U {D_\lambda}^{1/2}$$

로 구해지며 자명근에 해당하는 첫 열을 제외한 다음의 두 열을 이용하여 2차원 그래프를 그릴 수 있다. 다중대응분석에서는 이러한 그래프에서 좌표점의 상대적 위치를 가지고 데이터를 해석하게 된다. 허명회(1999)는 다중대응분석에서 2차원그래프에 대한 근사도를

$$GOA = \sum_{k=2}^{3} \eta_k / \sum_{k=2}^{l} \eta_k$$

로 정의한 바 있다. 여기서 $\eta_k = \lambda_k/(m-1), l = min(k_1 - 1, k_2 - 1)$이다.

두 유전자좌문제에서도 행렬도식 방법을 생각할 수 있다. 다음과 같이 중심화 및 표준화된 행렬

$$H^* = {D_s}^{-1/2}(B - rr')\,{D_s}^{-1/2} \tag{3.4}$$

를 생각한다. 이 때 r은 B의 대각원소로 이루어진 벡터이고 , $D_s = diag(s_1, \cdots s_{2(k_1+k_2)}), s_k = r_k/n^*(1 - r_k/n^*)$이다(Gower and Hand, 1996). 이를 분해한 것을

$$H^* = V D_\lambda V'$$

라 하고 좌표점으로

$$A^* = V {D_\lambda}^{1/2}$$

의 첫 두열을 사용한다. ${a_{1i}}^*$를 첫 유전자좌의 i번째 대립형질에 대한 수량화값, ${b_{2j}}^*$를 두번째 유전자좌의 j번째 대립형질에 대한 수량화값이라고 할때 (3.4)로부터 이들의 내적은

$${a_{1i}}^{*'} {b_{2j}}^* = \frac{p_{1i,2j} - p_{1i}p_{2j}}{\sqrt{p_{1i}(1-p_{1i})p_{2j}(1-p_{2j})}}$$

이 된다. 여기서 p_{1i}는 첫 유전자좌의 i번째 대립형질에 대한 관측비율이며, p_{2j}는 두 번째 유전자좌의 j번째 대립형질에 대한 관측비율을 의미한다. 또한 $p_{1i,2j}$는 첫 유전자좌에서 i번째 대립형질을 가지면서 두 번째 유전자좌의 j번째 대립형질을 가질 때의 관측비율이다. 따라서 이에 n^*배를 한 것이 첫 유전자좌의 i번째 범주에 속하냐 아니냐, 두번째 유전자좌의 j번째 범주에 속하냐 아니냐에 대한 2×2분할표의 카이제곱 검정통계량이 되어 두 범주가 독립인지를 알아볼 수 있게 된다. 편의상 이를 "다중대응행렬도"라고 하겠다. 행렬 H^*대신

$$H^{**} = {D_B}^{-1/2}(B - rr'){D_B}^{-1/2}$$

를 사용하면 이들의 내적이 전반적인 적합도 검정에 대한 카이제곱검정통계량과 관련이 있을 것이다. 다중대응분석은 실제로는 모든 데이터를 모두 2차원분할표로 생각하여 분석하는 방식이므로 여기서의 방법들은 모두 연관균형검정과 직접적인 관계를 갖지는 않는다. 만약 각 유전자좌가 하디-와인버그 평형을 만족한다면 각 유전자좌에서 대립형질의 조합으로 이루어진 범주들로 새로운 분할표, 즉 ${}_{k_1}C_2 \times {}_{k_2}C_2$의 이원분할표를 만들고 여기에 앞절에서의 단순대응분석 또는 단순대응행렬도를 적용함으로써 두 유전자좌가 독립인지를 파악할 수 있을 것이다.

4. 분석사례

여기서 사용되는 자료는 한국인 집단에서 서로 혈연관계가 없는 사람 492명을 추출한 것이다. 원자료는 17개의 유전자좌(D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, THO1, TPOX, CSF1PO, ACTBP2M, F13A1, FESFPS, D12S391)에 대해 빈도를 조사한 것이다(cf. 한길로 외, 1999). 여기서는 하나의 유전자좌에서 하디-와인버그 평형이 깨지는 유전자좌와 하디-와인버그 평형을 유지하는 유전자좌의 예로서 각각 F13A1와 D3S1358을 분석하였고, 두 유전자좌간의 관계를 알아보기 위한 방법의 예로 각각은 하디-와인버그평형을 만족하나 연관균형은 깨지는 유전자좌로서 vWA와 D8S1179를 분석하였다.

4.1. 하나의 유전자좌에서의 대립형질의 수량화와 그래프 결과

하나의 유전자좌에서의 문제를 위해 F13A1와 D3S1358을 분석하였다. 표 4.1과 표 4.2는 각각 F13A1와 D3S1358의 대립형질빈도 및 하디-와인버그 검정 결과로 F13A1의 경우 하디-와인버그평형이 깨지는 경우이고 D3S1358의 경우는 평형을 만족하는 경우의 예이다. 두 유전자좌에 대해 통상적인 대응분석을 적용하여 대립형질을 수량화한 결과 및 행렬도 방식의 수량화결과가 표 4.3과 표 4.4이다. 이 두 유전자좌에서는 행렬 G에 대해 각각 첫 번째와 두 번째 고유치가 음이 되어 행과 열의 좌표가 각각 x축, y축에 대응인 결과가 되었다. 그림 4.1과 그림 4.2는 F13A1의 그래프로 대응분석그래프 그림 4.1에서 비슷한 위치에 찍히는 대립형질은 유사한 행(열)프로파일을 갖는다는 것을 나타내는데 예를 들어 대립형질 3.2와 6은 비슷한 프로파일을 가지며 4와 5의 행프로파일은 서로 다르다. 대립형질 7은

표 4.1: F13A1의 빈도

allele	3.2	4	5	6	7
3.2	71				
4	46	9			
5	14	1	4		
6	165	51	20	109	
7	0	2	0	0	0
유전자비율	37.30	11.99	4.37	46.14	0.20
χ^2 검정	29.467 (p=0.001)				
정확검정	4.90E-10 (p=0.008)				

표 4.2: D3S1358의 빈도

allele	12	14	15	16	17	18	19
12	0						
14	0	0					
15	2	13	63				
16	2	13	121	38			
17	1	3	70	63	23		
18	0	2	32	25	16	1	
19	0	0	1	1	2	0	0
유전자비율	0.51	3.15	37.09	30.59	20.43	7.83	0.41
χ^2 검정	11.891 (p=0.9425)						
정확검정	1.68E-12 (p=0.885)						

다른 대립형질들과 매우 동떨어져 있어 아주 다른 행 프로파일을 가짐을 알 수 있다.이러한 결과는 표 4.5의 분할표에서 확인할 수 있다. 이 유전자좌는 하디-와인버그 평형이 깨지는 예로서 하디-와인버그 평형이 유지되는 D3S1358의 그래프 그림 4.3과 비교했을 때 좌표점이 중심으로부터 멀리 떨어져 찍히고 있음을 알 수 있다. 대응행렬도의 그림 4.2에서 행의 좌표는 직선으로 연결하고 열의 좌표는 연결하지 않았다. 이들 행과 열을 각각 하나씩 짝지어 해석할 수 있는데 우선 대립형질 7과 7, 7과 5는 거의 직각을 이루고 있어 독립임을 보여주는 반면 7과 4는 같은 방향으로 길게 누워있어 관측치가 기대치보다 크며 이 관계가 카이제곱통계량에 기여를 한다는 것을 알 수 있다. 5와 5도 역시 같은 방향으로 길게 있어 동형접합률이 높은 대립형질이라는 것을 알 수 있다. 4와 5의 경우는 둔각을 이루어 이들

의 관측치가 기대치보다 작은 방향으로 차이가 난다는 것을 알 수 있다. 이러한 결과는 모두 표 4.5의 결과와 일치한다. 또한 중심으로부터 많이 벗어난 대립형질 4, 5, 7이 하디-와인

표 4.3: F13A1 수량화 결과

	대응분석		대응행렬도분석			
	행수량화		행수량화		열수량화	
allele	1축	2축	1축	2축	1축	2축
3.2	0.014	-0.038	0.020	0.069	0.020	-0.069
4	0.306	0.202	0.242	-0.209	0.242	0.209
5	-0.675	0.056	-0.322	-0.035	-0.322	0.035
6	-0.034	-0.019	-0.053	0.040	-0.053	-0.040
7	1.592	-1.813	0.164	0.245	0.164	-0,245

표 4.4: D3S1358 수량화

	대응분석		대응행렬도분석			
	행수량화		행수량화		열수량화	
allele	1축	2축	1축	2축	1축	2축
12	-0.307	0.015	-0.070	0.004	0.070	0.004
14	-0.320	-0.181	-0.181	-0.113	0.181	-0.113
15	-0.020	-0.037	-0.038	-0.079	0.038	-0.079
16	0.123	-0.024	0.216	-0.046	-0.216	-0.046
17	-0.054	0.127	-0.077	0.202	0.077	0.202
18	-0.097	-0.024	-0.086	-0.023	0.086	-0.023
19	0.013	0.600	0.003	0.135	-0.003	0.135

버그 평형을 깨는 데 기여를 많이 하고 있다. 실제로 이들의 카이제곱통계량은 각각 8.82, 11.82, 7.33로서 전체 카이제곱통계량값 29.47의 94.9%를 차지한다. 여기서 1축과 2축의 고유값은 전체의 67.5%를 설명한다. D3S1358의 그래프 그림 4.3에서는 대립형질들이 대체로 원점주위에 몰려 있어 독립성가정에서 벗어나지 않고 있음을 시사한다. 또한 그림 4.4로부터 16과 16, 14와 17이 음의 방향으로. 16과 14가 양의 방향으로 관련이 있는 편임을 알 수 있다. 이 때 1축과 2축의 고유값은 전체의 82.3%이다.

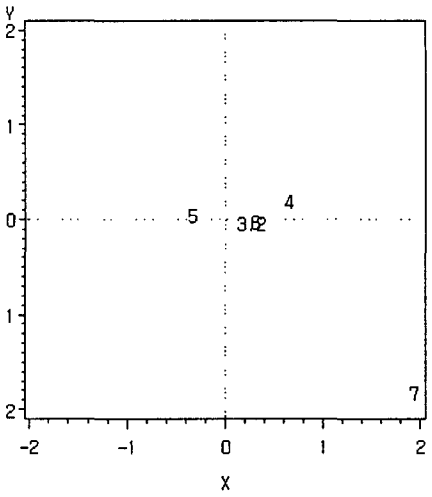

그림 4.1: F13A1 대응분석그래프

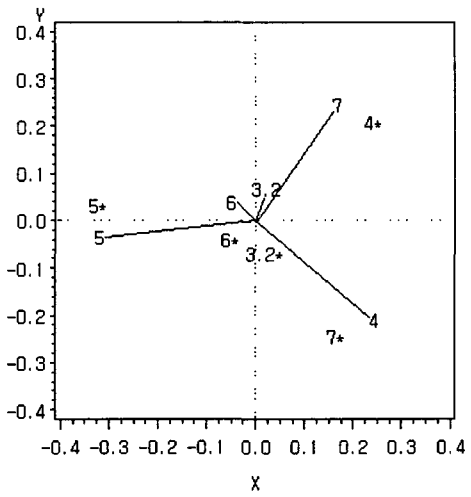

그림 4.2: F13A1 대응행렬도

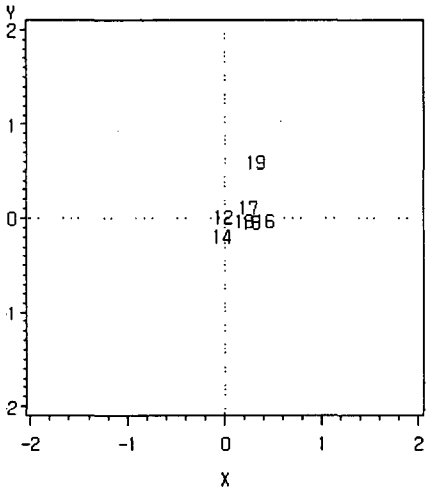

그림 4.3: D3S1358 대응분석그래프

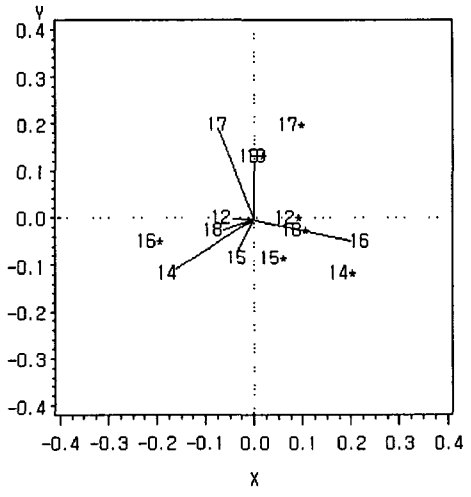

그림 4.4: D3S1358 대응행렬도그래프

표 4.5: F13A1의 분할표

각 칸의 값은 위로부터 관측수, 기대수, 셀카이제곱통계량, 행백분율을 나타낸다.

	3.2	4	5	6	7	total
3.2	71	23	7	82.5	0	0.71
	68.44	22.01	8.02	84.66	0.37	
	0.1	0.05	0.13	0.06	0.37	
	38.69	12.53	3.81	44.96	0.00	
4	23	9	0.5	25.5	1	8.82
	22.01	7.08	2.58	27.22	0.12	
	0.05	0.52	1.68	0.11	6.46	
	38.98	15.25	0.85	43.22	1.69	
5	7	0.5	4	10	0	11.82
	8.02	2.58	0.94	9.92	0.04	
	0.13	1.68	9.97	0.00	0.04	
	32.56	2.33	18.60	46.51	0.00	
6	82.5	25.5	10	109	0	0.8
	84.66	27.22	9.92	104.73	0.46	
	0.06	0.11	0.00	0.17	0.46	
	36.34	11.23	4.41	48.02	0.00	
7	0	1	0	0	0	7.33
	0.37	0.12	0.04	0.46	0.00	
	0.37	6.46	0.04	0.46	0.00	
	0.00	100.00	0.00	0.00	0.00	

표 4.6: D3S1358의 분할표

각 칸의 값은 위로부터 관측수, 기대수, 셀카이제곱통계량, 행백분율을 나타낸다.

	12	14	15	16	17	18	19	total
12	0	0	1	1	0.5	0	0	0.38
	0.01	0.08	0.93	0.76	0.51	0.2	0.01	
	0.01	0.08	0.01	0.07	0.00	0.2	0.01	
	0.00	0.00	40.0	40.0	20.0	0.00	0.00	
14	0	0	6.5	6.5	1.5	1	0	2.3
	0.08	0.49	5.75	4.74	3.17	1.21	0.06	
	0.08	0.49	0.1	0.65	0.88	0.04	0.06	
	0.00	0.00	41.94	41.94	9.68	6.45	0.00	
15	1	6.5	63	60.5	35	16	0.5	1.26
	0.93	5.75	67.7	55.83	37.28	14.28	0.74	
	0.01	0.1	0.33	0.39	0.14	0.21	0.08	
	0.55	3.56	34.52	33.15	19.18	8.77	0.27	
16	1	6.5	60.5	38	31.5	12.5	0.5	2.59
	0.76	4.74	55.83	46.04	30.74	11.78	0.61	
	0.07	0.65	0.39	1.40	0.02	0.04	0.02	
	0.66	4.32	40.20	25.25	20.93	8.31	0.33	
17	0.5	1.5	35	31.5	23	8	1	2.2
	0.51	3.17	37.28	30.74	20.53	7.86	0.41	
	0.00	0.88	0.14	0.02	0.3	0.00	0.86	
	0.50	1.49	34.83	31.34	22.89	7.96	1.00	
18	0	1	16	12.5	8	1	0	1.99
	0.2	1.21	14.28	11.78	7.86	3.01	0.16	
	0.2	0.04	0.21	0.04	0.00	1.34	0.16	
	0.00	6.45	8.77	8.31	7.96	2.6	0.00	
19	0	0	0.5	0.5	1	0	0	1.2
	0.01	0.06	0.74	0.61	0.41	0.16	0.01	
	0.01	0.06	0.08	0.02	0.86	0.16	0.01	
	0.00	0.00	0.27	0.33	1.0	0.00	0.00	

4.2. 두 개의 유전자좌에서의 대립형질의 수량화와 그래프 결과

두 유전자좌간의 관계를 위한 방법의 예로 vWA와 D8S1179를 분석하였다. vWA와 D8S1179는 동일 염색체상에 있지는 않으며, 개인식별을 위해 많이 사용되는 유전자좌이다(cf. Jorde et al., 2000; Han et al., 2000; Holt et al., 2000; Klintschar et al., 1999; Klintschar et al., 2001; Tracey,2001). 한국인 유전자자료에서 vWA와 D8S1179에 대한 정확검정의 p-값은 각각 0.696, 0.103로서 두 유전자좌는 하디-와인버그평형을 만족하나 연관균형에 대한 정확검정의 p-값은 0.025로 나타나 연관불균형이 발생하는 경우이다. 한편 vWA와 D8S1179의 분할표에 대한 카이제곱검정결과는 p-값이 0.383이다.

표 4.7은 두 유전자좌에 대해 얻어진 버트행렬에 들어가는 행렬로서 $Z_{11}{}'Z_{11}$ 과 $Z_{22}{}'Z_{22}$는 주변빈도를 보여주고 $Z_{11}{}'Z_{12}$과 $Z_{21}{}'Z_{22}$는 각각 vWA와 D8S1179에서의 대립형질들의 빈

도이며, $Z_{11}{}'Z_{21}$은 vWA와 D8S1179의 교차분할표이다. 여기서는 전체 돗수의 4배를 하는 대신 총합을 맞추고 그에 맞게 각 빈도를 1/2 또는 1/4로 조정하였다. 표 4.8와 표 4.9은 이에 대한 다중대응분석 및 행렬도방식의 수량화결과이다. 그림 4.5와 그림 4.6는 다중대응분석의 그래프인데 여기서 vWA의 21과 D8S1179의 8, 17이 중심에서 멀리 떨어져 있다. vWA 21과 D8S1179의 14가 대응되며 D8S1179의 17에는 vWA의 15등이 대응됨을 알 수 있다. 이 그래프의 근사도는 70.3%이다. 그림 4.7과 그림 4.8의 다중대응행렬도를 보면 vWA의 16과 D8S1179의 11, 15가 반대방향으로 길게 뻗어있어 기대값보다 관측수가 적은 방향으로 연관되어 있음을 알 수 있고, vWA의 17과 D8S1179의 16은 반대방향, vWA의 18과 D8S1179의 16은 순방향으로 연관이 되어 있다. 한편 하디-와인버그의 평형을 가정하고 각 유전자좌와의 조합을 다시 분할표로 작성하여 분석한 그림이 그림 4.9 - 그림 4.12에 있다. 그림 4.9과 그림 4.10은 대응분석을 이용한 분석결과로 vWA의 (15, 18)과 D8S1179의 (15, 17)등이 대응되며 vWA의 (15, 19)와 D8S1179의 (12, 15), (12, 17)등이 대응됨을 알 수 있다. 그림 4.11와 그림 4.12는 대응행렬도로서 vWA의 (14, 19)와 D8S1179의 (12, 15), (12, 17) , vWA의 (18, 18)과 D8S1179의 (10, 15)등이 내적이 커서 독립성가정에서 벗어나는데 기여함을 알 수 있다.

표 4.7: 버트 행렬을 구성하는 블럭행렬들

$Z_{11}{}'Z_{11}(vWA)$

	14	15	16	17	18	19	20	21
14	97.5	0	0	0	0	0	0	0
15	0	12.5	0	0	0	0	0	0
16	0	0	86.5	0	0	0	0	0
17	0	0	0	152	0	0	0	0
18	0	0	0	0	97.5	0	0	0
19	0	0	0	0	0	39.5	0	0
20	0	0	0	0	0	0	6	0
21	0	0	0	0	0	0	0	0.5

$Z_{11}{}'Z_{12}(vWA)$

	14	15	16	17	18	19	20	21
14	14	2	17	35.5	19.5	8.5	0.5	0.5
15	2	0	2	5.5	2.5	0.5	0	0
16	17	2	20	24	18	4.5	1	0
17	35.5	5.5	24	47	26.5	12.5	1	0
18	19.5	2.5	18	26.5	19	9.5	2.5	0
19	8.5	0.5	4.5	12.5	9.5	3	1	0
20	0.5	0	1	1	2.5	1	0	0
21	0.5	0	0	0	0	0	0	0

$Z_{11}{}'Z_{21}(vWA\ \&\ D8S1179)$

	8	10	11	12	13	14	15	16	17
14	0.25	11.25	8.75	16	24.25	15.75	13.75	6.5	1
15	0	0.75	1	2.5	3.5	1.75	2.25	0.5	0.25
16	0	12	5.25	13.75	23.25	15.5	11.75	5	0
17	0.5	15.75	16	19.5	38.25	27.25	23.75	9.5	1.5
18	0.5	13	8.5	15.25	20.75	16	15.25	8	0.25
19	0.25	2.75	3	6.5	9.5	7.5	5.5	4	0.5
20	0	0.75	0.5	1	1.5	1	1.25	0	0
21	0	0.25	0	0	0	0.25	0	0	0

$Z_{21}{}'Z_{22}(D8S1179)$

	8	10	11	12	13	14	15	16	17
8	0	0	0.5	0	0	0	1	0	0
10	0	6	7	8.5	13	12.5	5	4.5	0
11	0.5	7	4	6.5	5.5	9	8	2	0.5
12	0	8.5	6.5	17	17	11	8	5.5	1
13	0	13	5.5	17	34	23.5	20	7.5	0.5
14	0	12.5	9	11	23.5	15	10	4	0
15	1	5	8	8	20	10	14	6	1.5
16	0	4.5	2	5.5	7.5	4	6	4	0
17	0	0	0.5	1	0.5	0	1.5	0	0

$Z_{22}{}'Z_{22}(D8S1179)$

	8	10	11	12	13	14	15	16	17
8	0.5	0	0	0	0	0	0	0	0
10	0	135	0	0	0	0	0	0	0
11	0	0	65.5	0	0	0	0	0	0
12	0	0	0	62.5	0	0	0	0	0
13	0	0	0	0	119	0	0	0	0
14	0	0	0	0	0	88	0	0	0
15	0	0	0	0	0	0	18	0	0
16	0	0	0	0	0	0	0	2.5	0
17	0	0	0	0	0	0	0	0	1

표 4.8: vWA 부분 수량화 결과

allele	다중대응분석		다중대응행렬도분석	
	1축	2축	1축	2축
14	0.0400	-0.0551	-0.0069	0.0201
15	0.6914	-0.7859	0.0668	-0.0261
16	-0.8782	-0.1947	-0.3962	-0.326
17	0.3592	-0.3824	0.5525	-0.1285
18	-0.0800	0.6361	-0.2627	0.3466
19	0.4966	0.6237	0.0056	0.1308
20	-0.1835	0.6332	-0.0686	0.0528
21	-3.7940	-0.5572	-0.0479	0.0002

표 4.9: D8S1179 부분 수량화 결과

allele	다중대응분석		다중대응행렬도분석	
	1축	2축	1축	2축
8	3.8365	1.6883	0.0927	0.1536
10	-0.7985	0.1944	-0.2117	0.0317
11	0.6323	0.0077	0.1942	0.2241
12	-0.1691	0.4299	-0.2306	0.1868
13	-0.1735	-0.4699	0.0284	-0.5260
14	-0.3309	-0.2473	-0.0101	-0.1026
15	0.7888	-0.0510	0.2293	0.1747
16	0.1419	1.1576	-0.0562	0.1874
17	3.1874	-0.8647	0.1698	0.0830

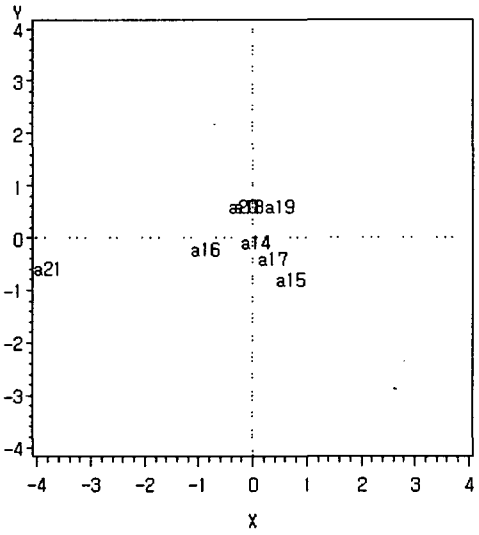

그림 4.5: 다중대응분석의 vWA부분

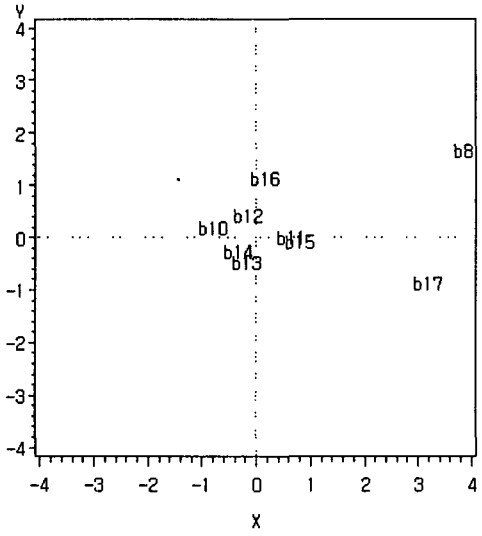

그림 4.6: 다중대응분석의 D8S1179부분

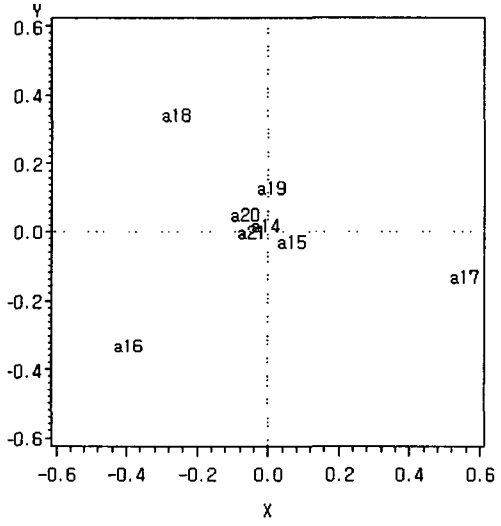

그림 4.7: 다중대응행렬도의 vWA부분

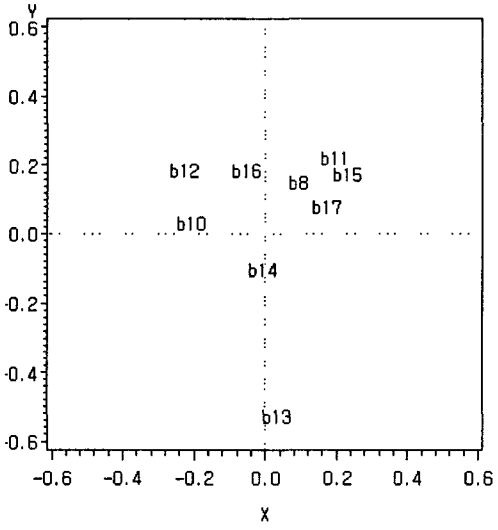

그림 4.8: 다중대응행렬도의 D8S1179부분

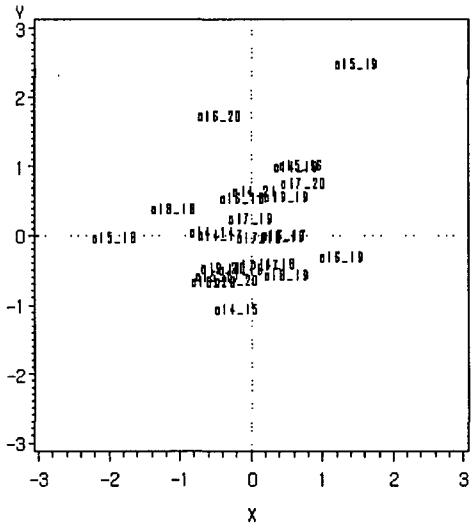

그림 4.9: 유전자조합의 대응분석vWA

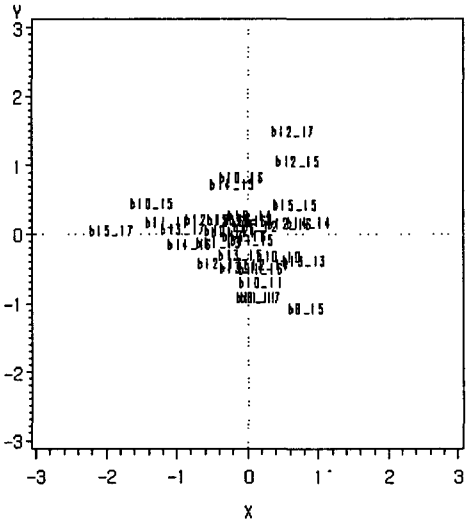

그림 4.10: 유전자조합 대응분석 D8S1179

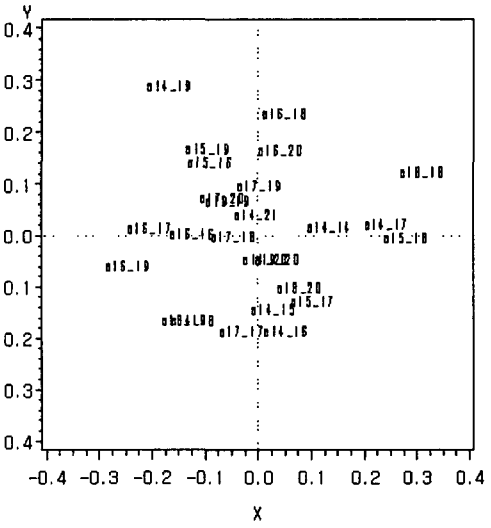

그림 4.11: 유전자조합의 행렬도 vWA

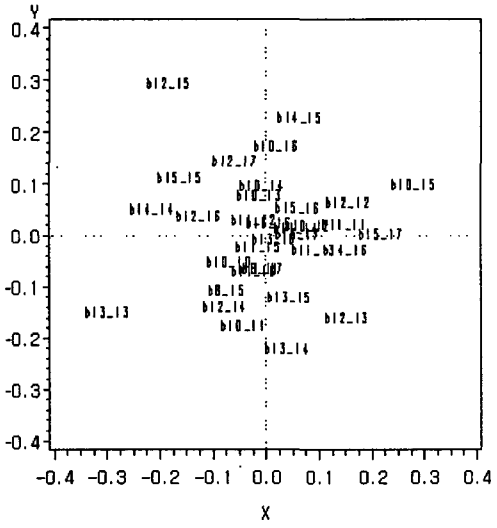

그림 4.12: 유전자조합의 행렬도 D8S1179

5. 결론 및 토의

본고에서는 유전자분석자료에서 각 유전자좌내의 대립형질 및 유전자좌간의 대립형질들간의 관계를 알아보기 위한 수량화 및 그래프화 방법으로서 대립형질간 거리를 표현할 수 있는 단순대응분석 및 다중대응분석과 거리보다는 내적에 의미가 있는 대응행렬도, 다중대응행렬도 등의 방법을 제시하였다. 다중대응분석의 경우에는 단순대응분석을 확장하는 과정에서 여러 가지 다른 형태의 접근방식이 있을 수 있는데(Krzanowski, 1995; Gower and Hand 1996), 여기서는 그 중에서 Park and Huh(1996)의 일반화정준상관분석과 연결되는 수량화방안을 사용하였다. 이 방법은 기하적인 해석에 근거한 것으로 여러 변수에 대해 보다 일관적으로 방법을 적용할 수 있다(허명회, 1999). 다중대응행렬도의 경우 에도 여러 가지 옵션이 있을 수 있겠으나 여기서는 내적이 개별적인 독립검정을 위한 카이제곱검정통계량을 나타낼 수 있도록 하는 방법을 적용하였다. 단순대응분석 및 대응행렬도를 사용했을 때의 결과는 유전자의 하디-와인버그 평형의 검정과 같은 결과를 보이나, 다중대응분석 및 다중대응행렬도를 사용하였을 때에는 연관불균형과 직접적으로 관계가 있지는 않다. 그 이유는 다중대응분석이 실제로는 다차원분석이 아니며 버트행렬이 단순히 2차원적인 분할표의 조합이기 때문이다. 따라서 하디-와인버그 평형이 유지된다는 가정하에서는 각 유전자좌에서의 유전자조합들을 가지고 다시 2원분할표를 만들어 단순대응 및 대응행렬도를 사용하는 것이 대안이 될 수 있다. 또한 로그-선형모형에 대한 행렬도 방식을 향후 연구할 수 있을 것이다. 분석된 그래프를 보면 주로 돗수가 적은 대립형질들이 원점에서 많이 벗어나 있는 모습을 볼 수 있다. 실제로 하디-와인버그 평형이나 연관균형이 깨지는 것이 주로 빈도가 높지 않은 대립유전자에서 발생하게 되는 특성이 있기는 하나 이러한 그래프를 통해 불균형을 일으키는 대립형질을 파악함으로써 몇 개의 관측치로 인한 섣부른 결론을 막을 수 있을 것이다. 근본적으로 카이제곱거리는 돗수가 작은 칸에 더 큰 비중을 주게 되어 돗수가 작은 칸이 전체를 잠식해 버리는 경향이 있으므로 대립형질간의 관계를 고려할 때 카이제곱거리외에 다른 거리를 정의하는 방법도 모색할 수 있을 것이다.

참고문헌

[1] 이재원, 박미라 (2000). 한국인 유전자형의 통계적 분석, <유전자자료프로필 구축 학술발표회>, 국립과학수사연구소.

[2] 최용석 (1993). <SAS 대응분석>, 자유아카데미,서울.

[3] 한길로, 이용욱, 이혜린, 김성민, 구태완, 강일호, 이혜승, 황적준 (1999). 한국인에서 9개 STR 유전좌의 대립유전자빈도 및 유전적 특성에 관한 연구, <한국법의학회지>, 23권 1호, 51-62.

[4] 허명회 (1998). <수량화 방법 I, II, III, IV>, 자유아카데미, 서울.

[5] 허명회 (1999). <다변량수량화>, 자유아카데미, 서울.

[6] 駒澤 勉 (1982). <數量化理論と デ-タ 處理>, 朝創書店, 東京.

[7] Alford, R.L., Hamond, H.A., Coto, I. and Daskey, C.T. (1994). Rapid and efficient Resolution of parentage by amplification of short tandem repeats. *American Journal of Human Genetics*, Vol. **55**, 190-195.

[8] Budowle, B., Nhari, L.Y., Moretti, T.R., Kanoyangwa, S.B., Masuka, E., Defenbaugyh, D.A. and Smerick, J.B. (1997). Zimbabwe black population data on six short tandem repeat loci -CSF1PO, TPOX, THO1, D3S1358, VWA and FGA, *Forensic Science Intrernational*, Vol. **90**, 215-221.

[9] Crow, J.F. (1988). Eighty years ago: the beginnings of population genetics, *Genetics*, Vol. **119(3)**, 473-6.

[10] Emigh, T.H. (1980). A comparison of tests for Hardy-Weinberg law, *Biometrics*, Vol. **36**, 627-642.

[11] Fregeau, C.J., Tan-Siew, W.F., Yap, K.H., Carmody, G.R., Chowm, S.T. and Fourney, R.M. (1998). Population genetic characteristics of the STR loci D21S11 and FGA in eight diverse human populations, *Human Biology*, Vol. **70(5)**, 813-44.

[12] Gabriel, K.R. (1971). The biplot graphics display of matrices with applications to principal component analysis, *Biometrika*, Vol. **58**, 453-467.

[13] Gehrig, C., Hochmeister, M., Borer, U.V., Dirnhofer, R. and Budowle, B. (1999). Swiss Caucasian population data for 13 STR loci using AmpFISTR profiler pplus and cofiler PCR amplification kits, *Journal of Forensic Science Medicine*, Vol. **44(5)**, 1035-1038.

[14] Geenacre, M. (1984). *Theory and applications of correspondence analysis*, Academic Press.

[15] Gower, J.C. and Hand, D.J. (1996). *Biplots*, Chapman and Hall.

[16] Greenacre, M. and Hastie, T. (1987). Geometric intepretation of corresponsence analysis, *Journal of the American Statistical Association*, Vol. **82**, 437-337.

[17] Guo, S.W. and Thomson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles, *Biometrics*, Vol. **48**, 361-372.

[18] Han, G.R., Lee, Y.W., Lee, H.L., Kim, S.M., Ku, T.W., Kang, I.H., Lee, H.S. and Hwang, J.J. (2000). A Korean population study of the nine STR loci FGA, vWA, D3S1358, D18S51, D21S11, D8S1179, D7S820, S13S317 and D5S818, *International Journal of Legal Medicine*, Vol. **114(1-2)**, 41-44.

[19] Hardy, G.H. (1908). Mendelian proportions in a mixed population, *Science*, Vol. **28**, 49-50.

[20] Holt, C.L., Staffer, C., Wallin, J.M., Lazaruk, K.D. Nguyen, T., Budowle, B. and Walsh, P.S. (2000). Practical applications of genotypic surveys for forensic STR testing, *Forensic Science International*, Vol. **112**, 91-109.

[21] Jorde, L.B., Shortsleeve, P.A., Henry, J.W., Vanburen, R.T., Hutchinson, L.E. and Rigley, T.M. (2000). Genetic analysis of the Utah population: a comparison of STR and VNTR loci, *Human Biology*, Vol. **72(6)**, 927-936.

[22] Klintschar, M. al-Hammndi, N, Reichenpfader, B. (1999). Population genetic studies on the tetrameric short random repeat loci D3S1358, VWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317 and D7S820 in Egypt, *Forensic Science International*, Vol. **104(1)**, 23-31.

[23] Klintschar, M., Al-Hammadi, N., Reichenpfader, B. (2001). Significant differences between Yemenite and Egyptian STR profiles and the influence on frequency estimations in Arabs, *International Journal of Legal Medicine*, Vol. **114**, 211-214.

[24] Krzanowski W.J. (1995). *Recent advances in descriptive multivariate analysis*, Oxford Science Publications.

[25] Lee, J.W., Lee, H.S., Park, M. and Hwang, J.J. (2001). Evaluation of DNA match probability in criminal case, *Forensic Science International*, Vol. **116**, 139-148.

[26] Park, M.R. and Huh, M.H. (1996). Quantification plots for several sets of variables, *Journal of the Korean Statistical Society*, Vol. **25(4)**, 599-601.

[27] Shoemaker, J., Painter, I. and Weir, B.S. (1998). A Baysian characterization of Hardy-Weinberg disequilibrium, *Genetics*, Vol. **149**, 2079-2088.

[28] Tracey, M. (2001). Short tandem repeat-based identification of individuals and parents, *Croatian Medical Journal*,Vol. **42(3)**, 233-238.

[29] Ward, R.H. and Sing, C.F. (1970). A consideration of the power of the chi-square test to detect inbreeding effects in natural populations, *American Naturalist*, Vol. **104**, 355-365.

[30] Zaykin, D., Zhivotovsky, L. and Weir, B.S. (1995). Exact tests for association between alleles at arbitrary numbers of loci, *Genetica*, Vol. **96**, 169-178.

[2001년 5월 접수, 2002년 1월 채택]

Quantification and Graphical Methods for DNA Fingerprinting*

Mira Park [1)]

ABSTRACT

To explore the relationships among frequencies for sets of alleles, within or between loci, is one of the first analyses in population genetic study. The general question is whether the frequency of a set of alleles is the same as the product of each of the separate allele frequencies. For two alleles of a single locus, Hardy-Weinberg equilibrium is tested and for an allele from each of two loci, linkage disequilibrium is tested. However, it is more useful if we can quantify and graphically represent this information. In this study, we suggest graphical methods to find associations between alleles. We also analyze the STR data of Korean population as an illustration.

Keywords: Hardy-Weinberg equilibrium; Linkage disequilibrium; Correspondence analysis; Multiple correspondence analysis; Correspondence biplot; Multiple correspondence biplot.

* This work was supported by Korea Research Foundation Grant(KRF-1999-003-D00072)

1) Assistant professor, Eulji University School of Medicine.
E-mail: mira@emc.eulji.ac.kr