응용통계연구 (2009)
22(2), 329–340

이환 형제 자료에 대한 유전적 연관성 분석 방법의 비교

고민진[1] · 임길섭[2] · 이학배[3] · 송기준[4]

[1]연세대학교 의학통계학과, [2]연세대학교 의학통계학과,
[3]연세대학교 응용통계학과, [4]연세대학교 의학통계학과

(2008년 10월 접수, 2008년 12월 채택)

요 약

질적 형질에 대한 유전적 연관성 분석은 크게 두 가지로 구분 할 수 있는데, 모형 기반 분석과 그렇지 않은 모형 무관 분석 방법이다. 복합질병의 경우 멘델의 유전법칙을 잘 따르지 않기 때문에 모형 기반 분석 방법을 사용하는 것보다 모형 무관 분석 방법을 사용하는 것이 효율적이라고 알려져 있다. 이러한 모형 무관 분석 방법 중 이환 형제 쌍 자료를 이용한 분석 방법은 형제 쌍 간의 유전적 일치 비율을 기준으로 공유하고 있는 대립유전자의 분포를 이용하는 것으로 크게 proportion test, mean test, minmax test로 구분 할 수 있다. 본 연구에서는 형제집단자료로 확장된 경우, 유전 형식에 상관없이 로버스트한 방법으로 알려진 minmax test에 형제 쌍의 가중치를 고려할 수 있는 방법들 즉, 동일 가중 방법, Suarez의 방법, Hodge의 방법, Sham 등의 방법을 적용하여 그 성능을 비교하였다. 모의실험 자료를 이용하여 비교한 결과 표식유전자의 빈도, 형질의 유전 형식, 형제수에 상관없이 Suarez의 방법이 가장 검정력이 높은 방법으로 드러났다. 또한, 동일 가중 방법을 제외하고는 표식유전자의 빈도가 높아질수록, 형제수가 많아질수록 더 높은 검정력을 보였고, 이러한 현상은 우성 유전 형식을 가정한 자료에서 더욱 두드러지게 나타났다.

주요용어: 연관성 분석, 이환 형제집단자료, 가중 방법, 검정력.

1. 서론

질병 이환의 유, 무와 같은 질적 형질(qualitative trait)에 대한 유전적 연관성분석(linkage analysis)은 염색체(chromosome)상에서 위치를 알고 있는 유전자인 표식유전자(marker)가 형질유전자와 매우 가까운 거리에 위치해 있기 때문에 감수분열 과정에서 유전자 재조합(recombination)이 일어나지 않고 자손에게 전해진다는 원리를 이용하여 관심 있는 형질유전자(trait)와 표식유전자의 연관성에 대해 검정하는 것이다 (Ott, 1999).

이러한 질적 형질에 관한 연관성 분석은 크게 두 가지로 구분 할 수 있는데, 전통적인 연관성 분석 방법으로 알려진 모형 기반 분석(model-based linkage analysis)과 그렇지 않은 모형 무관 분석(model-free linkage analysis) 방법이다.

모형 기반 분석 방법은 관심 있는 형질(trait)이 우성, 열성, 상호우성, 반성유전 등 멘델의 유전법칙을 따르는 무도병(huntington disease), 유방암(breast cancer), 낭포성섬유증(cystic fibrosis) 등의 질병일 경우 효율적이다. 그러나 당뇨(diabetes)나 심장병, 고혈압(hypertension), 정신분열증(schizophrenia)

이 논문은 2006년 정부재원(교육인적자원부 학술조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음 (KRF-2006-C00088).
[4]교신저자: (120-752) 서울시 서대문구 신촌동 134 연세대학교 의학통계학과, 연구조교수.
E-mail: biostat@yuhs.ac

등과 같은 복합질병(complex diseases)의 경우 멘델의 유전법칙을 잘 따르지 않기 때문에 모형 기반 분석 방법을 사용하는 것보다 모형 무관 분석 방법을 사용하는 것이 더 효율적이라고 알려져 있다 (Collins와 Morton, 1995; Kruglyak 등, 1996; Olson 등, 1999).

모형 무관 분석 방법은 관심 있는 형질에 대한 특별한 유전적 모형 즉, 형질의 유전 형식(mode of inheritance), 대립유전자의 빈도(allele frequency), 침투율(penetrance rate), 재조합률(recombination fraction) 등에 구애받지 않고 분석할 수 있다는 장점을 가지고 있다. 이러한 모형 무관 분석 방법은 형제 쌍(sib-pair) 자료를 이용하여 분석하는 방법을 그 시초로 하여 유전자 재조합에 대한 정보의 손실도 크지 않으면서 비용과 시간 면에서 훨씬 더 효율적인 이환 형제 쌍(ASP: affected sib-pair) 자료를 이용하는 방법으로 개선되어졌다 (Penrose 1935; Suarez 등, 1978; Thomson, 1986; Holmans, 1998). 이환 형제 쌍 자료를 이용한 분석 방법은 형제 쌍 간의 유전적 일치 비율(proportion of identity by descent)을 기준으로 공유하고 있는 대립 유전자의 분포를 이용하는 것으로 크게 proportion test, mean test, minmax test로 구분할 수 있다. 이러한 방법들은 오직 이환 형제 쌍 자료에만 적용 가능한 방법으로 이환 형제수가 3명 이상인 형제집단자료(sibship data)가 존재하는 경우에는 구성할 수 있는 각각의 형제 쌍 마다 가중치를 고려해야 하는데, 가중치를 고려한 연관성 분석 방법으로 mean test에 기초한 몇 가지 방법들 즉, 동일 가중 방법(All possible method), Suarez (Hodge와 Suarez, 1979)의 방법, Hodge (Hodge, 1984)방법, Sham (Sham 등, 1997)의 방법 등이 제시되었다. 아울러 이 방법들의 성능을 비교하는 다양한 연구가 진행되었다. 그런데, mean test는 유전 형식에 민감하게 영향을 받는 방법으로서 가중치를 고려하더라도 특정한 유전 형식을 따르는 경우 왜곡된 결과를 제공할 수 있는 한계를 가지고 있다 (Whittemore와 Tu, 1998).

본 논문에서는 이환 형제 쌍 자료에서 이환 형제수가 3명 이상인 형제집단자료로 확장된 경우, 이환 형제 쌍 자료를 이용하는 분석 방법 중 유전 형식에 상관없이 로버스트한 방법으로 알려진 minmax test (Whittemore와 Tu, 1998)에 기초하여 형제 쌍의 가중치를 고려할 수 있는 방법들을 논의하고 비교하고자 한다. 이때, 가중치를 주는 방법으로서 동일 가중 방법, Suarez의 방법, Hodge의 방법, Sham 등의 방법을 적용시킨다. 이상의 네 가지 방법을 비교하기 위해 실제자료의 틀에 기초한 모의실험자료를 생성하여 평가하고 또한 실제 고혈압 형제집단자료를 이용하여 평가한다.

2. 이환 형제 쌍 자료를 이용한 연관성 분석

2.1. 유전적 일치도

질적 형질에 대한 연관성 분석 특히, 모형 무관 분석 방법의 중요한 근간이 되는 유전적 일치도(identity by descent: 이하 IBD)는 한 쌍의 형제가 같은 부모로부터 동일한 대립유전자를 내려받는 것으로 관심 있는 형질이 멘델의 유전법칙을 따른다면, IBD한 대립유전자의 수가 0, 1, 2일 때 사전 확률은 $p_0 = 1/4$, $p_1 = 1/2$, $p_2 = 1/4$가 된다. 그러나, 형질이 멘델의 유전법칙을 잘 따르지 않는 경우에는 IBD한 확률분포를 추정해야 한다.

IBD한 확률분포를 추정하기 위해 몇 가지의 기본적인 가정을 다음과 같이 설정하여야 한다. 먼저, 관심 있는 형질이 두 개의 대립유전자(diallelic, (D, d))를 갖는데, 이때 관심 있는 형질의 민감도를 높이는 대립유전자를 D, 그렇지 않은 유전자를 d라고 표현하고 두 대립유전자의 빈도(frequency)를 각각 $p, q(1-p)$라고 하기로 한다. 또한, 특정한 유전형을 가지고 있을 때, 질병이 발생할 확률, 즉 침투율은 유전형(genotype)이 DD일 때 f_2, Dd는 f_1, dd는 f_0라고 정의한다. 아울러 관심 있는 형질 유전자와 표식유전자간의 유전자 재조합률은 θ라고 한다. 따라서, 이상의 가정에 따르면 유전적 모형은 5개의 모수(parameter)를 갖게 되고 이 모수들은 $\nu = (f_0, f_1, f_2, p, \theta)$인 벡터(vector)로 나타낼 수

있다. 이때의 유병률(population prevalence of the disease)은 $K_p = p^2 f_2 + 2pqf_1 + q^2 f_0$로 정의되며, 가법분산(additive variance)과 지배분산(dominance variance)은 $V_D = p^2q^2(f_2 - 2f_1 + f_0)^2$, $V_A = 2pq\{p(f_2 - f_1) + q(f_1 - f_0)\}^2$로 정의된다 (Suarez 등, 1978; Payami 등, 1984; Knapp과 Strauch, 2004). 부모의 유전형의 조합(mating type)을 $mt_{D'}$, 관심 있는 형질유전자의 IBD한 대립유전자의 수를 k라 하면, k개를 IBD한 대립유전자의 조건부확률(conditional probability)은

$$\begin{aligned} &P(k\text{개의 동일한 대립유전자를 갖는 확률}|\,\text{ASP}) \\ &= \frac{\sum_{mt_D} P(k\text{개의 동일한 대립유전자를 갖는 확률, ASP}|\,mt_D)P(mt_D)}{\sum_{mt_D} P(\text{ASP}|\,mt_D)P(mt_D)} \end{aligned} \tag{2.1}$$

로 정의된다. 즉,

$$\begin{aligned} P(2\text{개의 동일한 대립유전자를 갖는 확률}|\,\text{ASP}) &= \frac{K_P^2 + V_A + V_D}{4K_P^2 + 2V_A + V_D}, \\ P(1\text{개의 동일한 대립유전자를 갖는 확률}|\,\text{ASP}) &= \frac{K_P^2 + V_A}{4K_P^2 + 2V_A + V_D}, \\ P(0\text{개의 동일한 대립유전자를 갖는 확률}|\,\text{ASP}) &= \frac{K_P^2}{4K_P^2 + 2V_A + V_D} \end{aligned} \tag{2.2}$$

로 표현할 수 있다. 여기서 유전자 재조합률을 $\Psi = \theta^2 + (1-\theta)^2$로 나타낼 수 있으며 (Haseman과 Elston, 1972), 이 경우 유전자 재조합을 고려한 IBD의 조건부 확률분포는 식 (2.3)과 같다.

$$\begin{pmatrix} z_2(\nu) \\ z_1(\nu) \\ z_0(\nu) \end{pmatrix} = T \cdot \begin{pmatrix} P(2\text{개의 동일한 대립유전자를 갖는 확률}|\,\text{ASP}) \\ P(1\text{개의 동일한 대립유전자를 갖는 확률}|\,\text{ASP}) \\ P(0\text{개의 동일한 대립유전자를 갖는 확률}|\,\text{ASP}) \end{pmatrix},$$

$$T = \begin{pmatrix} \Psi^2 & 2\Psi(1-\Psi) & (1-\Psi)^2 \\ 2\Psi(1-\Psi) & 2(1-2\Psi-\Psi^2) & 2\Psi(1-\Psi) \\ 2(1-\Psi)^2 & 2\Psi(1-\Psi) & \Psi^2 \end{pmatrix}. \tag{2.3}$$

결과적으로 표식유전자에 대한 정보를 알고 있을 때, 그에 대한 분포는 식 (2.4)와 같다 (Suarez 등, 1978; Payami 등, 1984; Knapp와 Strauch, 2004).

$$\begin{aligned} z_2(\nu) &= \frac{K_P^2 + \Psi V_A + \Psi^2 V_D}{4K_P^2 + 2V_A + V_D}, \\ z_1(\nu) &= \frac{2K_P^2 + V_A + 2\Psi(1-\Psi)V_D}{4K_P^2 + 2V_A + V_D}, \\ z_0(\nu) &= \frac{K_P^2 + (1-\Psi)V_A + (1-\Psi)^2 V_D}{4K_P^2 + 2V_A + V_D}. \end{aligned} \tag{2.4}$$

2.2. 연관성 분석 방법

이환 형제 쌍 자료를 이용한 연관성 분석을 수행할 때 몇 가지 가정이 선행되어야 하는데, 첫 번째로 관심 있는 형질이 두 개의 대립유전자(D, d)를 가지며, 그 중 관심 있는 형질의 민감도를 높이는 대립유전자를 D로 간주한다. 두 번째로 부모의 유전형은 유전자 다형성(polymorphic)을 만족해야 한

다. 세 번째로, 부모간의 무작위 결합(random mating)을 만족하여야 하며, 네 번째로 하디와인버그 균형(Hardy-Weinbug equilibruim: HWE)상태를 만족해야하고, 마지막으로 연관불균형(linkage disequilibrium) 즉, 대립유전자가 D인 경우의 확률을 p라 할 때, d인 경우의 확률은 $1-p$가 된다고 가정한다 (Payami 등, 1984; Blackwelder와 Elston, 1985; Knapp 등, 1994a, 1994b; Knapp와 Strauch, 2004).

2.2.1. Mean test 독립적인 이환 형제 쌍 자료가 n개 존재하는 경우, 각 형제 쌍이 IBD한 대립유전자의 수가 각각 0, 1, 2일 때 그에 대한 i번째의 형제 쌍 자료의 확률분포는 $\hat{z}_{0i}$, $\hat{z}_{1i}$, $\hat{z}_{2i}$가 된다. mean test는 IBD한 대립유전자 수의 확률분포 즉, $0.5\hat{z}_{1i}+\hat{z}_{2i}$를 고려한 방법으로 관심 있는 형질유전자와 대립유전자가 연관되어 있지 않다는 귀무가설 하에 이 선형결합의 기대평균은 $E(\pi_i)=0.5z_{1i}+z_{2i}$가 된다. 즉, 검정통계량은

$$T=\frac{\sum_{i=1}^{n}(0.5\hat{z}_{1i}+\hat{z}_{2i})-\sum_{i=1}^{n}E(0.5z_{1i}+z_{2i})}{\sqrt{\sum_{i=1}^{n}\mathrm{Var}\,(0.5\hat{z}_{1i}+\hat{z}_{2i})}} \tag{2.5}$$

로 자유도가 $n-1$인 t분포를 따르게 되며, 우성 유전(dominant inheritance)인 경우 검정력(power)이 뛰어나다고 알려져 있다 (Blackwelder과 Elston, 1985; Olson 등, 1999).

2.2.2. Proportion test proportion test는 IBD한 대립유전자가 2개인 확률분포 즉, $\hat{z}_{2i}$만을 고려한 방법으로 관심 있는 형질유전자와 대립유전자가 연관되어 있지 않다는 귀무가설 하에 이 선형결합의 기대평균은 $E(\pi_i)=z_{2i}$가 된다. 즉, 검정통계량은

$$T=\frac{\sum_{i=1}^{n}\hat{z}_{2i}-\sum_{i=1}^{n}E(z_{2i})}{\sqrt{\sum_{i=1}^{n}\mathrm{Var}\,(\hat{z}_{2i})}} \tag{2.6}$$

로 자유도가 $n-1$인 t분포를 따르게 되며, 열성 유전(recessive inheritance)인 경우 검정력이 뛰어나다고 알려져 있다 (Blackwelder과 Elston, 1985; Olson 등, 2002).

2.2.3. Minmax test minmax test에 대해 설명하기에 앞서, IBD한 대립유전자의 확률분포를

$$F_a=\left[z(\lambda):z=\lambda(0,a,1-a)+(1-\lambda)\left(\frac{1}{4},\frac{1}{2},\frac{1}{4}\right);0\le\lambda\le1\right] \tag{2.7}$$

로 재정의 한다. 여기서, λ는 실수(scalar), a는 고정된 상수(fixed constant)라 할 때, IBD한 대립유전자의 확률분포에 대한 선형결합, $v_0z_0+v_1z_1+v_2z_2$의 계수 중, v_1은 $\{(1/2)a\}/(1-a)\,(0\le a\le 1/2)$로 나타 낼 수 있으며, 이 경우의 검정통계량은 식 (2.8)로 정의된다.

$$T=\frac{\sqrt{n}\{4(a-1)\hat{z}_0+(6a-4)\hat{z}_1+3-4a\}}{\sqrt{3-8a+6a^2}}, \tag{2.8}$$

여기서 관심 있는 형질의 실제 IBD한 대립유전자의 확률분포를 따르는 경우 검정통계량은 식 (2.9)이다.

$$T = \frac{\sqrt{n}\{3 - 4a + 4(a-1)z_0 + (6a-4)z_1\}}{\sqrt{3 - 8a + 6a^2}} \tag{2.9}$$

이때, 어떤 고정된 상수 a에 대해 a_*의 볼록함수(convex function)를 $f(a, a_*)$라 한다. 이 볼록함수의 끝점(end point)이 mean test을 나타내는 값 $a_* = 0.5$와 proportion test를 나타내는 $a_* = 0$인 경우, 최대페널티(maximum penalty)는 $\max_{0 \le a_* \le 0.25} f(a, a_*)$이다.

이 두 개의 볼록함수는 $a' = (3 - \sqrt{6})/(4 - \sqrt{6}) = 0.355$에서 만난다. 즉, minmax test의 a'값을 결정하는 방법은 $\max_{0 \le a_* \le 0.25} f(a', a_*) = \min_{0 \le a \le 0.25}[\max_{0 \le a_* \le 0.25} f(a, a_*)]$와 같다. 여기서, 최대 패널티 값이 가장 작아지는 값은 두 볼록함수가 만나는 $a' = 0.355$로 이 값은 위에서 정의한 수식에 의해 $v_1 = 0.275$가 된다.

결국, IBD한 대립유전자의 확률분포에 대한 선형결합은 $\hat{\pi}_i = 0.275\hat{z}_{1i} + \hat{z}_{2i}$이며, 이때의 검정통계량은

$$T = \frac{\sum_{i=1}^{n}(0.275\hat{z}_{1i} + \hat{z}_{2i}) - \sum_{i=1}^{n} E(0.275 z_{1i} + z_{2i})}{\sqrt{\sum_{i=1}^{n} \mathrm{Var}\,(0.275\hat{z}_{1i} + \hat{z}_{2i})}} \tag{2.10}$$

로 자유도가 $n-1$인 t분포를 따르게 된다.

minmax test는 우성형질에서 검정력이 뛰어난 mean test와 열성형질에서 검정력이 뛰어난 proportion test에 의해 얻어지는 것으로 유전형식에 상관없이 로버스트한 방법으로 3장에서는 이런 minmax test를 기초로 이환 형제수가 3명 이상인 형제집단자료를 분석하는 방법에 대해 논의한다.

3. 가중을 고려한 연관성 분석 방법

3.1. 동일 가중 방법

이환 형제수가 3명 이상인 형제집단자료에서 i번째 가족의 형제 수를 n_i라 하면, 총 $\{n_i(n_i-1)\}/2 = N$쌍이 된다. 이처럼, 동일 가중 방법은 모든 가능한 형제 쌍의 조합들을 독립적인 쌍으로 고려한 방법으로 검정통계량은

$$T = \frac{\sum_{i=1}^{N} \pi_i - \sum_{i=1}^{N} E(\pi_i)}{\sqrt{\sum_{i=1}^{N} \mathrm{Var}(\pi_i)}} \tag{3.1}$$

로 자유도가 $N-1$인 t분포를 따르게 된다. 여기서 minmax test에 기초하였으므로 IBD한 대립유전자의 확률분포에 대한 선형결합은 $\pi_i = 0.275 z_{1i} + z_{2i}$이다. 이 방법은 동일한 가족 내에서 구성된 형제 쌍들을 서로 독립적인 관계로 설정한다는 점에서 다소 왜곡된 추론을 할 수 있다는 단점을 가지고 있다.

3.2. Suarez의 방법

이환 형제수가 3명 이상인 형제집단자료에서 i번째 가족의 형제 수를 n_i라 하면, 총 $\{n_i(n_i-1)\}/2 = N$쌍이 되는데, 이중 $n_i - 1$쌍만이 통계학적으로 독립이 된다. 만약, 모든 가능한 형제 쌍의 조합 중 통

표 **3.1.** 형제가 3명인 경우 결합분포와 조건부분포

			$R(1,3)$			
	2		1		0	
$R(1,2)$	$R(2,3)$		$R(2,3)$		$R(2,3)$	
2	2	1/16 (1)	2	0 (0)	2	0 (0)
	1	0 (0)	1	2/16 (1)	1	0 (0)
	0	0 (0)	0	0 (0)	0	1/16 (1)
1	2	0 (0)	2	2/16 (1/2)	2	0 (0)
	1	2/16 (1)	1	0 (0)	1	2/16 (1)
	0	0 (0)	0	2/16 (1/2)	0	0 (0)
0	2	0 (0)	2	0 (0)	2	1/16 (1)
	1	0 (0)	1	2/16 (1)	1	0 (0)
	0	1/16 (1)	0	0 (0)	0	0 (0)

계학적으로 독립인 $n_i - 1$쌍만을 고려한다면, 형제 쌍에 대한 유전적 정보를 모두 고려하지 못하므로, $\{n_i(n_i-1)\}/2$쌍에 대해 $2/n_i$의 가중치를 고려한다. 이 방법은 결과적으로 $n_i - 1$쌍만 고려하는 경우로 생각 할 수 있다. 결국, 동일 가중 방법에 $2/n_i$의 가중치를 고려한 경우로 검정 통계량은

$$T = \frac{\sum_{i=1}^{N} \frac{2}{n_i}\pi_i - \sum_{i=1}^{N} \frac{2}{n_i}E(\pi_i)}{\sqrt{\sum_{i=1}^{N}\left(\frac{2}{n_i}\right)^2 \mathrm{Var}(\pi_i)}} \tag{3.2}$$

로 자유도가 $N-1$인 t분포를 따르게 된다. 여기서도 IBD한 대립유전자의 확률분포에 대한 선형결합은 $\pi_i = 0.275z_{1i} + z_{2i}$이다.

3.3. Hodge의 방법

Hodge의 방법은 부모의 유전형의 조합이 $ab \times cd$로 자식이 가질 수 있는 유전형이 ac, ad, bc, bd인 경우를 가정한다. 이에 대해 이환 형제수가 3명인 형제집단자료는 $R(1,2)$, $R(1,3)$, $R(2,3)$의 세 가지 형제 쌍을 가지며, 첫 번째 형제 쌍을 ac라 하면 두 번째, 세 번째 형제 쌍은 네 가지의 유전형을 가질 수 있는데, 세 번째 형제 쌍의 경우는 표 3.1에서처럼 독립이 아니다.

첫 번째와 두 번째 형제 쌍은 서로 독립이며, 세 번째 형제 쌍은 다른 두 쌍의 조건에 의해 결정된다는 중요한 사실을 표 3.1을 통해 확인 할 수 있다. 이에 대해 이환 형제수가 4명 이상인 형제집단자료인 경우도 이환 형제수가 3명인 형제 집단자료처럼 첫 번째 형제 쌍, 두 번째 형제 쌍에서 n_i번째 형제 쌍 인 $n_i - 1$ 형제 쌍은 $R(1,2), R(1,3), \ldots, R(1,n_i)$이고, 두 번째 형제 쌍, 세 번째 형제 쌍에서 n_i번째 형제 쌍인 $n_i - 2$ 형제 쌍은 $R(2,3), \ldots, R(2,n_i)$이다.

결국, $n_i(n_i-1)/2$번째 형제 쌍은 $n_i - 1$, n_i번째 형제 쌍을 나타내는 것이다.

k개의 범주를 가지며, 각 범주가 $p_1, \ldots, p_k$의 확률을 갖는 경우 다항 분포(multinomial distribution)를 따를 때, 엔트로피(entropy) 또는 선택적 정보(selective information)는 $H = \sum_i p_i \log_2(p_i)$로 정의된다. 만약, X, Y의 두개의 확률변수를 고려한다면, p_{ij}는 $P(X = i, Y = j)$의 결합 확률이고 $p_{i\cdot}$은 $P(X = i)$의 주변 확률(marginal probability), $\sum_j p_{ij}$는 $P_{j|i}$인 조건부 확률이다.

결국, X가 주어졌을 때 Y의 조건부 정보(conditional information)는 $H(Y|X) = -\sum_i p_{i\cdot} \sum_j p_{j|i}$

$\log_2(p_{j|i})$이다. 이때, 이환 형제수가 3명 이상인 형제집단자료의 총 정보(total information), Shannon의 정보는 $H_{n_i} = 2n_i - 3 + 0.5^{n_i-1}$로 정의되며, 순서척도에 따라 이환 형제수가 3명~10명인 형제집단자료는 총 정보를 3/2로 값, $H = \{2(2n_i - 3 + 0.5^{n_i-1})\}/3$이 독립인 형제 쌍에 대한 동일한 정보량이 된다. 결국 이환 형제수가 n_i명인 형제집단자료의 경우 $H_i = \{4(2n_i - 3 + 0.5^{n_i-1})\}\{3n_i(n_i-1)\}$의 가중치를 고려하게 된다. 결과적으로 Hodge의 방법의 검정통계량은

$$T = \frac{\sum_{i=1}^{N} H_i\pi_i - \sum_{i=1}^{N} H_i E(\pi_i)}{\sqrt{\sum_{i=1}^{N} H_i^2 \mathrm{Var}(\pi_i)}} \tag{3.3}$$

로 자유도가 $N-1$인 t분포를 따르게 된다. 여기서도 IBD한 대립유전자의 확률분포에 대한 선형결합은 $\pi_i = 0.275z_{1i} + z_{2i}$이다.

3.4. Sham 등의 방법

Sham 등의 방법은 질병에 이환된 형제와 그렇지 않은 형제를 모두 포함하는 형제집단을 고려하는데, n개의 형제집단에서 $a_1, \ldots, a_n$명의 이환 형제와 $u_1, \ldots, u_n$명의 이환되지 않은 형제가 존재한다고 가정한다. a_i명의 이환 형제를 가진 i번째 형제집단에서 이환 형제 쌍은 $a_i(a_i-1)/2$가지가 존재할 수 있으므로 전체 자료에서 형제 쌍의 수는 $\sum a_i(a_i-1)/2$가지가 된다. Y_{ijk}를 i번째 형제집단의 이환된 형제 j와 k간의 IBD한 대립유전자의 수라고 하면 i번째 형제집단 전체 IBD 값은 $Y_i = \sum Y_{ijk}$로 표현할 수 있다. 귀무 가설하에서 Y_i의 평균과 분산을 각각 m_i와 s_i^2이라고 한다면 검정통계량은 다음과 같이 나타낼 수 있다.

$$T = \frac{\sum w_i Y_i - \sum w_i m_i}{\sqrt{\sum w_i^2 s_i^2}} \tag{3.4}$$

이 통계량은 근사적으로 표준정규분포를 따르는데, 이 때 w_i는 i번째 형제집단에 대한 가중치로서 다음과 같은 값을 갖는다. 여기에서 μ_i는 대립 가설하에서 Y_i의 평균을 말한다.

$$w_i = \frac{\left(\frac{\mu_i - m_i}{s_i^2}\right)}{\left(\frac{\mu_1 - m_1}{s_1^2}\right)}. \tag{3.5}$$

4. 모의실험자료를 이용한 비교

이환 형제가 3명 이상인 경우의 형제자료에 대해 minmax test을 기초로 한 동일 가중 방법, Suarez의 방법, Hodge의 방법, Sham 등의 방법의 검정력을 평가하기 위해 모의실험자료를 생성하였다.

모의실험자료는 재조합률을 $\theta = 0.01$로 관심 있는 형질유전자와 표식유전자가 연관되어 있다는 가정 아래, 유병률 $K_P = 22.45\%$로서 한국인의 고혈압 유병율을 적용하였다 (질병관리본부, 2003). 또한, 표식유전자의 빈도 0.1, 0.3, 0.5에 대해 유전 형식을 우성 유전과 열성 유전으로 구분하여 그림 4.1과 같이 생성하였다. 이러한 기본적인 가정 아래 이환 형제가 3, 4, 5명으로 구성된 핵가족을 각각 100번씩 반복하여 총 1,800가족, 7,200명의 형제집단자료를 생성하였다.

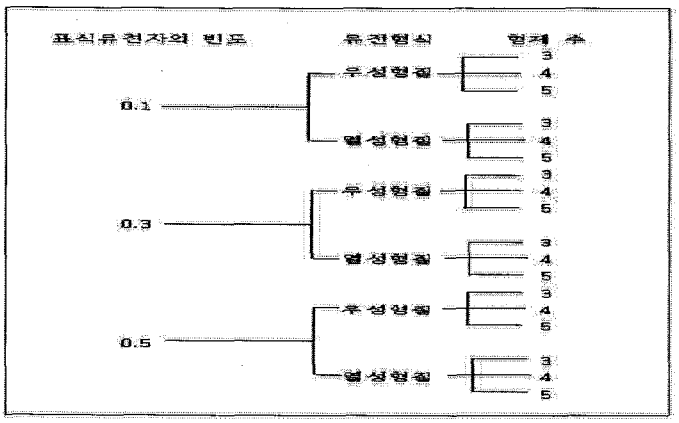

그림 **4.1.** 모의실험자료의 구성

표 **4.1.** 모의실험 결과

Marker	Sib size	Dominant(mean, SD)				Recessive (mean, SD)			
		All	Suarez	Hodge	Sham	All	Suarez	Hodge	Sham
0.1	3	20 (0.18, 0.13)	48 (0.12, 0.13)	47 (0.13, 0.13)	48 (0.11, 0.12)	14 (0.25, 0.15)	29 (0.19, 0.16)	28 (0.20, 0.16)	29 (0.20, 0.17)
	4	31 (0.16, 0.15)	75 (0.08, 0.14)	69 (0.09, 0.14)	73 (0.08, 0.13)	14 (0.23, 0.16)	49 (0.14, 0.16)	44 (0.15, 0.16)	44 (0.15, 0.16)
	5	33 (0.12, 0.09)	89 (0.02, 0.03)	82 (0.03, 0.04)	84 (0.02, 0.03)	11 (0.23, 0.14)	57 (0.10, 0.13)	50 (0.12, 0.14)	53 (0.11, 0.12)
0.3	3	19 (0.18, 0.12)	49 (0.11, 0.11)	43 (0.12, 0.11)	44 (0.11, 0.12)	11 (0.22, 0.13)	33 (0.15, 0.13)	30 (0.16, 0.13)	31 (0.15, 0.13)
	4	35 (0.12, 0.11)	79 (0.03, 0.06)	72 (0.05, 0.07)	77 (0.03, 0.07)	15 (0.20, 0.12)	54 (0.08, 0.09)	46 (0.10, 0.10)	49 (0.08, 0.09)
	5	47 (0.08, 0.07)	100 (0.01, 0.01)	97 (0.01, 0.02)	99 (0.01, 0.01)	16 (0.21, 0.13)	63 (0.07, 0.10)	54 (0.09, 0.11)	60 (0.08, 0.10)
0.5	3	20 (0.17, 0.13)	49 (0.11, 0.13)	46 (0.12, 0.13)	48 (0.12, 0.12)	17 (0.21, 0.14)	37 (0.15, 0.14)	34 (0.16, 0.14)	33 (0.15, 0.15)
	4	36 (0.11, 0.11)	89 (0.03, 0.09)	86 (0.04, 0.09)	88 (0.03, 0.08)	14 (0.21, 0.13)	59 (0.10, 0.14)	50 (0.12, 0.14)	55 (0.11, 0.13)
	5	50 (0.07, 0.06)	100 (0.01, 0.01)	100 (0.01, 0.01)	100 (0.01, 0.01)	17 (0.16, 0.09)	82 (0.03, 0.04)	71 (0.04, 0.06)	76 (0.04, 0.05)

18가지 경우에 대한 모의실험자료의 생성은 SIMLA(SIMulation of Linkage and Association) 3.1(http://wwwchg.duhs.duke.edu)를 이용하였다. SIMLA는 관심 있는 형질과 표식유전자의 재조합률, 관심 있는 형질의 유병률, 관심 있는 형질의 유전 형식, 표식유전자의 빈도, 가족 내의 형제 수를 지정했을 때, 모의실험자료를 생성할 수 있는 프로그램이다.

또한 모의실험자료의 IBD한 대립유전자의 확률분포를 추정하기 위해서 본 연구에서는 S.A.G.E.(Statistical Analysis for Genetic Epidemiology) 5.1(http://darwin.cwru.edu)의 서브루틴 중 하나인 GENIBD을 이용하여 IBD 분포를 추정하였다. 이러한 정보들을 이용하여 각 방법의 검정통계량 값을 구하는 것은 R(http://www.r-project.org)을 이용하였다.

표 **5.1.** 표식유전자의 빈도

Chromosome	Marker	Allele(frequency)
1	AGT2	A(0.22)
		G(0.78)
	AGT3	A(0.96)
		C(0.04)
	AGT4	A(0.71)
		G(0.29)
	AGT5	C(0.73)
		T(0.27)
17	ACE6	D(0.30)
		I(0.70)
19	ApoE112	C(0.22)
		T(0.78)
	ApoE158	C(0.96)
		T(0.04)

18가지 경우에 대한 모의실험자료를 100번 반복하여 분석하였을 때, 세 가지 방법에 대하여 유의확률(p-value)을 구하였고 아울러 이 유의확률이 유의수준(significance level) 0.05이하가 되는 즉, 귀무가설이 기각되는 빈도(%)를 구하여 비교하였다.

표식유전자의 빈도와 형제 수에 따라 세 가지 방법에 대한 유의확률의 평균과 표준편차 및 0.05이하가 되는 빈도를 표 4.1을 통해 제시하였다. 표식유전자의 빈도가 증가함에 따라 네 가지 방법 모두에서 유의확률이 낮아지는 경향을 보였는데, 좀 더 구체적으로 살펴보면 형질의 유전 형식과 형제 수에 상관없이 Suarez방법의 결과가 동일 가중 방법, Hodge의 방법, Sham 등의 방법보다 유의확률이 더 작은 값으로 얻어졌다. 이러한 경향은 유의확률이 0.05이하가 되는 빈도에서도 동일하게 나타났다. 또한, 네 가지 방법 모두 우성 유전인 경우 유의확률이 낮게 나오는 경향을 보였다. 동일 가중 방법의 경우, 유전 형식에 상관없이, 형제수가 증가하여도 유의확률 값이나 빈도에서 그 변화 양상에 뚜렷한 차이를 보이지 않았다. 그러나, Suarez의 방법, Hodge의 방법, Sham 등의 방법에서는 형제수가 증가할수록 유의확률이 더 작게 얻어졌다.

5. 실제자료를 이용한 분석

연세대학교 심혈관질환 유전체 연구센터에서 조사된 유전체자료를 이용하여 이환 형제수가 2명 이상인 경우 minmax test를 기초로 동일 가중 방법, Suarez의 방법, Hodge의 방법, Sham 등의 방법을 비교하였다.

분석에 사용된 CGC자료는 자식들이 모두 고혈압에 걸린 핵가족(nuclear family)을 대상으로 형제가 2명으로 구성된 가족은 11가족, 3명으로 구성된 가족은 3가족, 4명으로 구성된 가족은 2가족으로 총 16가족, 39명으로 구성되었다. 표식 유전자는 AGT(Angiotensinogen)에서 유전소(locus) 즉, G-217T(AGT2), A-20C(AGT3), G-6A(AGT4), M235T(AGT5)와 ACE(Angiotensin Converting Enzyme)에서 유전소 ALU(ACE6), ApoE(ApolipoproteinE)에서 유전소 A3932G(ApoE112), G4070A(ApoE158)을 이용하였다. 이때 표식유전자들의 빈도는 표 5.1과 같다.

일곱 개의 표식유전자를 이용하여 동일 가중 방법, Suarez의 방법, Hodge의 방법을 수행했을 때 결과는 표 5.2와 같다. 유의수준을 0.05라 했을 때, 고혈압과 표식유전자가 연관되어 있지 않다는 귀무가설 하

표 5.2. 실제자료 분석결과

Marker	P-value(Power)			
	All	Suarez	Hodge	Sham
AGT2	0.124(0.472)	0.221(0.345)	0.202(0.365)	0.198(0.370)
AGT3	0.696(0.109)	0.713(0.105)	0.710(0.106)	0.701(0.108)
AGT4	0.058(0.621)	0.101(0.514)	0.092(0.533)	0.097(0.522)
AGT5	0.089(0.541)	0.150(0.431)	0.138(0.449)	0.123(0.473)
ACE6	0.392(0.220)	0.491(0.174)	0.477(0.180)	0.425(0.203)
ApoE112	0.001(0.969)	0.002(0.949)	0.001(0.969)	0.001(0.969)
ApoE158	0.747(0.097)	0.806(0.085)	0.799(0.087)	0.768(0.093)

에 ApoE112가 세 가지 방법 모두에서 고혈압유전자와 연관되어 있다는 결과를 보였다. ApoE112를 제외한 다른 표식유전자의 경우 귀무가설 하에 유의한 결과를 보이지는 않았지만, 모든 표식유전자에서 세 가지 방법 중 동일 가중 방법이 Suarez의 방법, Hodge의 방법, Sham 등의 방법보다 유의확률이 다소 낮게 얻어졌으나 큰 차이는 없었고, 이는 검정력을 계산했을 때에도 비슷한 양상을 보였다.

6. 결론 및 고찰

지금까지 이환 형제집단자료에 가중치를 부여하고 이를 이용하여 유전적 연관성을 검정하는 방법들 즉, 동일 가중 방법, Suarez의 방법, Hodge의 방법, Sham 등의 방법을 설명하고 그 성능을 비교하였다. 본 연구에서 비교된 방법들은 형질의 유전 형식에 상관없이 로버스트하다고 알려진 minmax test에 기초한 것들로서 이는 IBD한 대립유전자의 확률분포에 대한 선형결합, $\pi_i = 0.275z_{1i} + z_{2i}$를 적용하는 것이다.

본 연구에서는 이 네 가지 방법을 비교하고, 이들 중 유전적 연관성을 판단하는데 가장 검정력이 높은 방법을 찾고자 하였다. 모의실험자료를 이용하여 비교한 결과 표식유전자의 빈도, 형질의 유전 형식, 형제 수에 상관없이 Suarez의 방법이 가장 검정력이 높은 방법으로 드러났다. 또한, 동일 가중 방법을 제외하고는 표식유전자의 빈도가 높아질수록, 형제수가 많아질수록 더 높은 검정력을 보였다. 이러한 현상은 열성 유전보다 우성 유전을 가정한 자료에서 더욱 두드러지게 나타났다.

기존의 연구들 중 mean test에 기초하여 각각의 형제 쌍마다 가중치를 고려한 방법들을 비교한 결과에서는 동일 가중 방법이 높은 검정력을 보였다 (Sham 등, 1997). 이것은 본 연구에서 Suarez의 방법이 높은 검정력을 보인 것과는 다소 다른 양상을 나타낸 것이다. 그러나 mean test의 경우 우성 유전을 하는 형질에서 검정력이 뛰어난 방법으로 유전 형식을 정확히 알 수 없는 대부분의 복합형질에서는 본 연구에서 제시한 minmax test에 기초한 Suarez의 방법을 사용하는 것이 타당하다고 판단된다.

본 연구에서는 질병에 이환된 형제집단자료에서 형제 쌍들의 가중치를 고려한 연관성 분석 방법들을 논의하고 비교하였는데, 이를 기초로 몇 가지의 확장된 접근이 가능하다고 여겨진다. 먼저, 형제집단을 구성하고 있는 형제들 중 질병에 이환되지 않은 구성원이 존재할 경우, 이들의 유전적 정보를 포함시켜 형제 쌍에 대한 가중치를 고려하는 것이다. 이는 이환된 형제들만의 정보를 이용하여 IBD한 대립유전자의 수를 추정하는 것보다 더 정확한 추정치를 얻는데 도움을 줄 수 있기 때문이다. 즉, 질병에 이환되지 않은 형제들을 포함시켜 형제 쌍을 구성한 후, 그에 따른 가중치를 주어 연관성 분석을 수행하는 방법의 개발이 필요할 것으로 판단된다.

다음으로 형제집단자료의 범위를 확장시켜 일반 가계(general pedigree)자료에서 이환된 친척의 쌍(relative pair)을 이용하여 연관성 분석을 할 경우, 친척의 쌍에 대한 가중치를 효율적 고려할 수 있는 연

구가 이루어져야 할 것이다. 실제 유전적 연관성을 파악하기 위한 각종 연구들에서 핵가족뿐만 아니라 3세대 이상의 일반 가계자료를 조사하는 것이 대부분이므로 그때 존재 할 수 있는 이환된 친척 쌍을 고려하여 연관성 분석을 수행할 필요가 있는데, 이 경우 특정한 친척 쌍마다 적절하게 가중치를 고려하는 방법은 좀 더 정확한 분석 결과를 얻는데 많은 기여를 할 것으로 생각된다.

참고문헌

질병관리본부 만성병조사팀 (2003). 우리나라 고혈압 유병률 및 관리현황, 국민건강영양조사 3기 검진조사 보고서.

Blackwelder, W. C. and Elston, R. C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci, *Genetic Epidemiology*, **2**, 85–97.

Collins, A. and Morton, N. E. (1995). Nonparametric tests for linkage with dependent sib pairs, *Human Heredity*, **45**, 311–318.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics*, **2**, 3–19.

Hodge, S. E. (1984). The information contained in multiple sibling pairs, *Genetic Epidemiology*, **1**, 109–122.

Hodge, S. E. and Suarez, B. K. (1979). A simple method to detect linkage for rare recessive disease: An application to juvenile diabetes, *Clinical Genetics*, **15**, 126–136.

Holmans, P. (1998). Affected sib-pair methods for detecting linkage to Dichotomous traits: Review of the methodology, *Human Biology*, **70**, 1025–1040.

Knapp, M., Seuchter, S. A. and Baur, M. P. (1994a). Two-locus disease models with two marker loci: The power of affected-sib-pair tests, *American Journal of Human Genetics*, **55**, 1030–1041.

Knapp, M., Seuchter, S. A. and Baur, M. P. (1994b). Linkage analysis in nuclear families. 1: Optimality criteria for affected sib-pair tests, *Human Heredity*, **44**, 37–43.

Knapp, M. and Strauch, K. (2004). Affected-sib-pair test for linkage based on constraints for identical-by-Descent distributions corresponding to disease models with imprinting, *Genetic Epidemiology*, **26**, 273–285.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach, *American Journal of Human Genetics*, **58**, 1347–1363.

Olson, J. M., Goddard, K. A. and Dudek, D. M. (2002). A second locus for very-late-onset Alzheimer disease: a genome scan reveals linkage to 20p and epistasis between 20p and the amyloid precursor protein region, *American Journal of Human Genetics*, **71**, 154–161.

Olson, J. M., Witte, J. S. and Elston, R. C. (1999). Tutorial in biostatistics genetic mapping of complex traits, *Statistics in Medicine*, **18**, 2961–2981.

Ott, J. (1999). *Analysis of Human Genetic Linkage*, 3rd edition, The johns Hopkins University Press, Baltimore.

Payami, H., Thomson, G. and Louis, E. J. (1984). The affected sib method. Ⅲ. Selection and recombination, *American Journal of Human Genetics*, **36**, 352–362.

Penrose, L. S. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage, *Annals of Engenics*, **6**, 133–138.

Sham, P. C., Zhao, J. H. and Curtis, D. (1997). Optimal weighting scheme for affected sib-pair analysis of sibship data, *Annals of Human Genetics*, **61**, 61–69.

Suarez, B. K., Rice, J. and Reich, T. (1978). The generalized sib pair IBD distribution: Its use in the detection of linkage, *Annals of Human Genetics*, **42**, 87–94.

Thomson, G. (1986). Determining the mode of inheritance of RFLP-associated diseases using the affected sib-pair method, *American Journal of Human Genetics*, **39**, 207–221.

Whittemore, A. S. and Tu, I. P. (1998). Simple, Robust linkage tests for affected sibs, *American Journal of Human Genetics*, **62**, 1228–1242.

Comparison of Methods for Linkage Analysis of Affected Sibship Data

Go Min Jin[1] · Kil Seob Lim[2] · Hak Bae Lee[3] · Kijun Song[4]

[1]Department of Biostatistics, Yonsei University; [2]Department of Biostatistics, Yonsei University; [3]Department of Applied Statistics, Yonsei University; [4]Department of Biostatistics, Yonsei University

(Received October 2008; accepted December 2008)

Abstract

For complex diseases such as diabetes, hypertension, it is believed that model-free methods might work better because they do not require a precise knowledge of the mode of inheritance controlling the disease trait. This is done by estimating the sharing probabilities that a pair shares zero, one, or two alleles identical by descent(IBD) and has some specific branches of test procedure, *i.e.*, the mean test, the proportion test, and the minmax test. Among them, the minmax test is known to be more robust than others regardless of genetic mode of inheritance in current use. In this study, we compared the power of the methods which are based on minmax test and considering weighting schemes for sib-pairs to analyze sibship data. In simulation result, we found that the method based on Suarez' was more powerful than any others without respect to marker allele frequency, genetic mode of inheritance, sibship size. Also, The power of both Suarez- and Hodge-based methods was higher when marker allele frequency and sibship size were higher, and this result was remarkable in dominant mode of inheritance especially.

Keywords: Linkage analysis, affected sibship data, weighted method, power.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD, Basic Research Promotion Fund)(KRF-2006-C00088).

[4]Corresponding author: Research Assistant Professor, Department of Biostatistics, Yonsei University, Seoul 120-749, Korea. E-mail: biostat@yuhs.ac