선형모형에 기초한 질적 형질과 일배체의 관련성 분석방법 비교

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

한 혜 리

선형모형에 기초한 질적 형질과 일배체의 관련성 분석방법 비교

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2005년 6월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
한 혜 리

한혜리의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2005년 6월 일

차 례

그 림 차 례

표 차 례

국 문 요 약

선형모형에 기초한 질적 형질과 일배체의 관련성 분석 방법 비교

질병과 관련 있는 유전자를 찾기 위한 최근의 방법 중의 하나는 인간의 유전자에 존재하는 단일염기다형성(single nucleotide polymorphism; 이하 SNP)을 이용하는 것이다. 그러나 많은 질병들이 복합형질(complex trait)이므로 단일 SNP을 이용하는 방법에는 제약점이 있다. 이러한 한계를 극복하기 위해 여러 SNP들을 이용하여 일배체(haplotype)를 구축하는 방법이 제안되었다. 특히 SNP들이 독립적이지 않고 연관불균형(linkage disequilibrium)상태일 경우 일배체로 군집화하여 질병과의 관련성을 보는 것이 효과적인 방법이다.

일배체를 구성함에 있어서 가장 중요한 문제는 일배체가 어떻게 구성되어 있는지를 알 수 없다는 사실이다. 따라서 일배체 빈도를 추정하기 위한 방법이 필요한데, 가장 대표적인 방법으로 EM 알고리즘을 들 수 있다. 그러나 EM 알고리즘은 안정적인 성능에도 불구하고 많은 SNP들이 존재하는 경우에는 추정이 불가능하다는 단점이 존재하여 이를 응용한 알고리즘들이 제시되었고 ECM 알고리즘, EE 알고리즘이 그것이다.

구축된 일배체와 질병과의 관련성을 보기 위하여 본 논문에서는 선형모형에 기초한 네 가지 분석방법을 비교하였다. 결측치 유무에 따라 완전자료와 불완전자료로 나누어서 비교하였으며 환경요인을 고려한 경우와 그렇지 않은 경우도 비교하였다. 네 가지 방법을 실재자료를 이용하여 분석한 결과 여러 특징들을 볼 수 있었는데 일배체를 구성하는 SNP수가 많아질수록 각 방법간의 성능차이가 관찰되었다. HTR(haplotype trend regression)의 경우는 SNP수와 상관없이 일관되지 못

한 결과를 보여주었고 Chaplin 방법이 가장 보수적인 방법인 것으로 보였으며 가장 덜 보수적인 방법은 스코어 방법인 것으로 나타났다. 또한 Chaplin 방법은 SNP수가 많은 경우 적은 결측치에도 결과가 많이 달라지는 것으로 나타났다. HWE 가정을 따르지 않는 자료인 경우는 스코어 방법이나 H-plus 방법을 사용하는 것이 바람직한 것으로 보였다.

환경요인을 고려한 경우는 그렇지 않은 경우에 비해 질적 형질과 일배체의 관련성이 떨어지는 것으로 보아 환경요인을 고려하지 않고 관련성 분석을 할 경우 형질과 일배체의 관련성을 과대 해석할 수 있다는 것을 알 수 있었고 방법간의 비교 측면에서는 스코어 방법과 H-plus 방법간의 차이가 대체적으로 없는 것으로 보인다.

핵심되는 말 : 일배체, 관련성분석, 선형모형, HTR 방법, 스코어 방법, Chaplin 방법, H-plus 방법

제 1장 서 론

1.1 연구 배경

인간 유전체지도(Human Genome Map)의 완성은 포스트 유전체(post Genome) 분야에 대한 연구의 시작을 의미하였으며(International Human Genome Sequencing Consortium, 2001), 최근의 연구들로 인간 개개인의 변이(variation)의 대부분은 DNA 염기 서열의 변이에 기인하고 이 중 많은 부분이 염기 치환(substitution)에 의한 것임이 알려졌다. 이러한 단일 염기쌍에서의 염기 치환을 SNP이라고 하며 인간의 유전자에서 10%이상의 빈도를 보이는 SNP의 개수는 5,300,000개로 추정되었다(Kruglyak and Nickerson, 2001).

질병과 관련 있는 유전자를 찾는 가장 일반적인 방법은 환자-대조군의 실험 설계를 통해 혈연관계가 없는 개인(unrelated individuals)들로부터 단일 표식유전자(marker)와 질병 정보를 수집하여 질병과 SNP의 관련성을 분석하는 것인데, 단일 SNP 정보를 이용하여 각각의 SNP과 질병의 직접적인 관련성을 분석하거나 여러 개의 SNP들로 일배체(haplotype)를 구성하여 이 일배체와 질병간의 관계를 분석하는 방법이 있다.

단일 SNP를 이용한 분석방법과 일배체를 이용한 분석방법은 장단점이 존재한다. 질병과 관련이 있다고 알려진 SNP들의 수 자체가 작다고 한다면 모든 일배체 조합을 고려한 분석보다 단일 SNP을 이용한 분석이 보다 효과적임이 알려져 있다. 그러나 SNP들이 연관불균형 상태에 있을 때는, 일배체를 이용한 관련성 분석이 단일 SNP를 이용한 분석보다 효과적임이 알려져 있으며(Akey and Xiong, 2001), 유전소(locus)에 여러 개의 대립유전자(allele)가 있을 경우 단일 SNP와 일배체를 이용한 분석 모두 검정력을 잃게 되나, 일배체를 이용한 분석이 덜 잃게 된다(Slager et al., 2000). 또한 일배체를 이용한 분석은 자유도가 증가하는 것을 피하면서 차원(dimension)을 줄일 수 있어서 관련성 분석을 할 때 검정력을 높여

주는 이점을 가지며 유전자를 판독(genotyping)하는 노력이나 비용적인 측면에서도 절감의 효과를 가진다.

전통적인 카이제곱검정방법으로부터 선형모형에 기반을 둔 회귀모형, 로지스틱 회귀모형, 일반화선형모형 등의 방법들, 일배체를 군집화(clustering)하는 방법에 이르기까지 최근 몇 년 동안 일배체와 질병의 관련성을 분석하는 방법들이 제시되었다. 이 중 선형모형에 기반을 둔 응용방법들이 제안되고 있는데 이는 선형모형을 이용한 방법들이 장점들을 지니고 있기 때문이다. 첫째, 선형모형은 유전요인이 아닌 성별이나 나이 등의 환경요인들을 고려할 수 있고 둘째, 특정 일배체의 효과를 수치로 표현할 수 있다는 것이다. 셋째, 일배체와 환경요인의 교호작용을 모델화 할 수 있으며, 넷째, 선형모형은 전통적으로 많이 개발되어온 모형이어서 응용방법이 많다는 점이다.

1.2 연구 목적 및 내용

국내에서 선형모형으로 일배체와 질병의 관련성을 제시한 바가 있으며(이은혜, 2004), 양적형질과(quantitative trait)과 일배체의 관련성을 단순선형모형과 일반화선형모형 두 가지를 이용해서 비교하였다. 본 논문에서는 기존의 연구 결과를 확장시켜 환자-대조군 자료를 이용하여 양적형질이 아닌 질병유무라는 질적형질(qualitative trait)과 일배체의 관련성 분석 방법을 비교하고자 한다. 질병이나 기형 등은 이분형 혹은 범주형으로 나누어질 수 있기 때문에 질적형질을 통한 분석방법은 인간의 유전자를 이용한 질병의 원인규명에 있어서 근간이 된다 할 수 있기 때문이다. 또한 기존의 연구결과는 형질과 일배체의 관련성을 보는데 있어서 유전적 요인만을 고려하였다. 그러나 특정 형질은 유전요인만이 아니라 환경요인과도 관련을 가지고 있는 경우가 많기 때문에 본 논문에서는 환경요인을 고려한 분석방법도 고려하고자 한다.

위의 두 가지 상황을 고려하여 HTR 방법(haplotype trend regression; Zaykin

et al., 2002)과 스코어 방법(Schaid et al., 2002), Chaplin 방법(Epstein and Satten et al., 2003), H-plus 방법(Zhao et al., 2003)의 관련성 분석 방법을 선택하였다. 관련성 분석은 수집한 유전체자료를 이용하여 일배체를 구축하고, 구축된 일배체와 형질과의 관련성을 분석하는 단계로 구성되는데, 첫 번째 단계에서는 HTR 방법과 스코어 방법은 EM 알고리즘을 사용하며, Chaplin 방법은 ECM 알고리즘, H-plus 방법은 EE 알고리즘을 사용하여 일배체의 빈도를 추정하여 일배체를 구축한다. 두 번째 단계에서는 구축된 일배체와 형질과의 관련성을 선형모형에 근거하여 분석하는데, HTR 방법은 단순선형모형, 스코어 방법은 일반화선형모형, Chaplin과 H-plus는 로지스틱회귀모형을 사용한다. 이상의 네 가지 방법으로 심혈관계유전체연구센터에서 수집한 유전체자료를 이용하여 일배체와 질적형질의 관련성을 분석하는 방법을 비교하고자 한다.

본 논문의 구성은 다음과 같다. 제 1장에서는 연구배경과 목적을 제시하고 제 2장에서 일배체 빈도 추정 방법에 대해 설명한다. 선형모형에 기초한 관련성 분석 방법에 대해서 이론적인 내용을 제 3장에서 소개하고 제 4장에서는 실재 환자-대조군 자료를 이용하여 분석방법을 비교한다. 마지막으로 제 5장에서 결론 및 논의할 점에 대해 말한다.

제 2장 일배체 빈도 추정 방법

2.1 개요

일배체란 하나의 염색체 상에서 가깝게 연관되어 있는 대립유전자들의 조합으로 정의될 수 있는데 재조합(recombination)이나 반복적 돌연변이(recurrent mutation)가 일어나지 않는다면 두개의 대립유전자(bi-allele)에 존재하는 SNP의 수가 N개 일 때 유전학적으로 구성될 수 있는 일배체의 수는 N+1보다 크지 않을 것으로 예측되나, 이론적으로는 SNP간의 자유로운 재조합을 가정한다면 2^N개이다 (Patil et al., 2001).

일배체를 이용한 분석에 있어서 가장 큰 제약점은 일배체가 어떻게 구성되어 있는지를 정확히 알 수 없다는 것이며 이를 phase를 알 수 없다라고 표현한다. 단일분자희석(single-molecule dilution)방법이나(Stephens et al., 1990), 긴 범위를 이용한 대립유전자 PCR(long-range allele-specific PCR)같은 방법(Micahlatos-Beloin et al., 1996), 이배체-반수체 전환(diploid-to-haploid conversion)방법(Douglas et al., 2001)이 지금까지 알려져 있는 일배체의 구성을 알기 위한 실험적인 접근방법들이다. 그러나 이러한 방법들은 비용이 많이 들고 중요한 기술적 문제들이 미해결 상태이어서 널리 사용되어지고 있지 못하다. 따라서 통계적인 방법으로 일배체의 빈도를 추정하는 여러 알고리즘들이 제안되었고 위의 실험적인 방법에 비해 정확하고 비용측면에서도 효과적인 해결책으로 생각되어 지고 있다.

혈연관계가 없는 개인 자료에서 일배체의 빈도를 추정하는 통계적 방법으로는 다음의 몇 가지 알고리즘이 존재한다. 첫 번째 방법은 일배체를 구축하는 알고리즘 중에서 초기에 제안되었던 방법으로서 최대절약성의 원리(principle of maximum parsimony)에 기반을 두고 있는 클락 알고리즘(Clark, 1990)이다. Clark은 모든 일배체가 최대절약성의 원리에 따라 추정되어 진다면 정확한 유일해를

갖게 된다고 말한다. 두 번째 방법은 EM 알고리즘(EM algorithm)을 이용한 방법으로서 표본의 일배체 빈도를 최대 우도(maximum likelihood)에 근거하여 추정한다. 이밖에도 유사깁스표집기(pseudo-Gibbs sampler)를 이용한 알고리즘(Stephens et al., 2001)과 긴 범위의 일배체의 구성을 알기 위하여 분할-정복(divide and conquer)의 알고리즘을 사용하는 PL(partition-ligation) 방법이 있다. 이 방법은 긴 일배체를 작은 단위의 결합이라고 생각하기 때문에 많은 수의 유전소가 존재할 때 유용하다. 사전 확률을 통계 모형에 결합하여 일배체 빈도를 추정하는 베이지안 방법(Bayesian method)도 제안되었다(Stephens et al., 2001). 제 2장에서는 일배체 빈도를 추정하는 알고리즘 중에서 EM 알고리즘과 그의 응용방법인 ECM 알고리즘, EE 알고리즘에 대해 논의하기로 한다.

2.2 EM 알고리즘

2.2.1 EM 알고리즘 개요

EM(expectation-maximization) 알고리즘(Dempster et al., 1977)은 결측값이 존재하는 불완전자료(incomplete data)를 다루기 위한 방법으로 불완전자료에서 완전자료(complete data)를 다루는 기법을 적용하여 모수를 추정하고 주어진 모델에 근거하여 결측값을 대체하는 것이다. EM 알고리즘은 E-step(estimation-step)과 M-step(maximization-step)이라고 불리는 2가지 단계를 반복적으로 수행함으로써 불완전자료(incomplete data)를 다루게 된다.

Excoffier와 Slatkin은 EM 알고리즘을 이용한 일배체 추정 방법을 최초로 제안하였는데(Excoffier and Slatkin, 1995), SNP들의 HWE(Hardy-Weinbug equilibrium)을 가정하며 최대우도를 이용하여 관찰된 데이터의 확률을 최적화하는 일배체의 확률을 추정한다.

2.2.2 EM 알고리즘 절차

EM 알고리즘에서 우도함수(likelihood function)는

$$L(\Theta) = P(G \mid \Theta) = \prod_{i=1}^{n} \sum_{(a,b) : a \oplus b = g_i} \theta_a \theta_b$$

이다. 위의 우도함수에서 G는 n개의 사람에서 관찰된 phase를 알 수 없는 유전형 자료이고 g_i는 i번째 사람에서 관찰된 phase를 알 수 없는 유전형 자료, Θ는 전체 일배체 빈도, θ_a와 θ_b는 각각 일배체 a와 b의 일배체 빈도, $a \oplus b = g_i$는 일배체 쌍 (a,b)가 i번째 관측된 유전형 자료 g_i를 구성한다는 것을 의미한다. HWE 가정하에서 EM 알고리즘은 반복적인 과정을 통하여

$$\theta_a^{(k+1)} = E_{\Theta^{(k)}}(n_a \mid G) / 2n$$

을 추정한다. 여기서 $\Theta^{(k)}$는 일배체 빈도의 현재 추정치(estimate)이고, n_a는 G에서 일배체 a의 개수이다. EM 알고리즘에서 주의할 점은 실제의 모집단의 빈도와 근접하게 일배체의 빈도에 대한 초기값을 정하는 것이 중요하다는 점이다.

EM 알고리즘의 장점은 EM 알고리즘의 성능이 HWE 가정의 위배에 크게 영향을 받지 않는다는 것이다. 특히 동형접합(homozygosity)을 많이 보이는 경우에도 EM 알고리즘의 성능은 모의실험을 통해서 HWE 가정의 위배 여부에 크게 영향을 받지 않는 것으로 연구결과 밝혀졌다(Niu et al., 2002). EM 알고리즘의 단점은 EM 알고리즘의 성능이 초기값 Θ의 영향을 많이 받는다는 것이다. 또한 국소적 최대값(local maxima)이 존재하는 경우 국소적 최적 최대 우도 추정값(locally optimal maximum likelihood estimates)으로 수렴하여 일배체의 개수가 많은 경우 문제가 될 수 있다.

2.3 ECM 알고리즘

2.3.1 ECM 알고리즘 개요

결측값이 없는 완전한 자료일 경우에도 모수가 많은 모델은 최대우도 추정값(maximum likelihood estimates)을 찾는 것이 쉽지 않다. 그러나 모델의 전체 모수 중에서 일부분을 알고 있을 때 나머지 모수를 추정하는 것은 전체 모수를 동시에 추정하는 것보다는 쉬운 일이 될 것이다. 이렇게 완전한 자료에서 모수들을 여러 개로 분할한 다음 하나의 집합을 제외한 나머지 집합을 조건부로 하나의 집합의 모수를 추정하는 방법을 CM(conditional Maximization) 알고리즘이라고 한다(Meng and Rubin, 1993).

ECM 알고리즘은 EM 알고리즘과 CM 알고리즘의 장점을 조합한 알고리즘이다(Meng and Rubin, 1993). ECM 알고리즘은 EM 알고리즘의 M-step(maximization step)을 두 개의 CM-step(conditional maximization step)으로 대체한 방법인데, 이렇게 함으로써 전체 모수를 한번에 추정하는데서 생기는 EM 알고리즘의 문제점을 전체모수를 여러 모수 집합으로 나누어 추정하는 방법으로 해결하고 있다. 또한 ECM 알고리즘은 우도의 단조수렴성이라는 측면에서 EM 알고리즘이 갖고 있는 안정성(stability)도 갖고 있다. ECM 알고리즘의 구체적인 방법은 다음에서 논의한다.

2.3.2 ECM 알고리즘 절차

E-step에서는 주어진 모수 추정값과 관측된 유전형 자료를 통해 결측되어 있는 일배체를 대치한다. 첫 번째 CM-step에서는 β를 갱신(update)하고 두 번째 CM-step에서는 p를 갱신한다. EM 알고리즘에서는 β와 p를 동시에 갱신하였다.

ECM 알고리즘에서 phase가 알려져 있는 경우 우도함수는 아래와 같다.

$$L_{FULL} = \prod_{(h,h')} \pi_{hh'}^{c_{hh'}} \rho_{hh'}^{d_{hh'}} = \frac{(\prod_{(h,h')} e^{(X_{hh'}^T \beta) d_{hh'}})(\prod_h P_h^{m_h})}{(\prod_{(h,h')} e^{X_{hh'}^T \beta} P_h P_{h'})^d}$$

위의 식에서 $c_{hh'}$과 $d_{hh'}$은 일배체 쌍(h,h')를 가지고 있는 대조군과 환자군의 수를 의미하고, m_h는 일배체 h의 개수를 의미한다.

E-step

$(k+1)$번째 단계를 생각해 보자. k번째 단계에서 β와 p에 대한 추정값인 $\beta^{(k)}$와 $p^{(k)}$를 얻었으므로 (k+1)단계에서 일배체 쌍(h,h')를 가지고 있는 대조군의 수를 추정할 수 있으며 식은 아래와 같다.

$$c_{hh'}^{(k+1)} = \sum_g c_g \frac{p_h^{(k)} p_{h'}^{(k)} I\{(h,h') \in S(g)\}}{\sum_{(h_1,h_2) \in S(g)} p_{h_1}^{(k)} p_{h_2}^{(k)}}$$

같은 방식으로 환자군의 수도

$$d_{hh'}^{(k+1)} = \sum_g d_g \frac{e^{X_{hh'}^T \beta^{(k)}} p_h^{(k)} p_{h'}^{(k)} I\{(h,h') \in S(g)\}}{\sum_{(h1,h2) \in S(g)} e_{h_1 h_2}^{T\ \beta^{(k)}} p_{h_1}^{(k)} p_{h_2}^{(k)}}$$

으로 추정할 수 있으며, 위 식에서 일배체 쌍이 유전형과 같을 때는

$I\{(h,h') \in S(g)\}$=1 이고 그렇지 않은 경우에는 0이다.

CM-step : update β

$\{d_{hh'}^{(k+1)}\}$와 $\{p_h^{(k)}\}$를 이용하면 로그우도의 최대화를 통해서 $\beta^{(k+1)}$을 갱신할 수 있다. 로그우도는 β에 의존하며 식은 아래와 같다.

$$log\,(L_\beta^{(k+1)}) \propto \sum_{(h,h')} d_{hh'}^{(k+1)} X_{hh'}^T \beta - d log\,(\sum_{(h,h')} e^{X_{hh'}^T \beta} p_h^{(k)} p_{h'}^{(k)})$$

여기서 준-뉴튼(quasi-Newton) 알고리즘을 이용하여 로그우도를 최대화할 수 있는데, 준-뉴튼 알고리즘을 이용하면 β의 스코어 식을

$$\frac{d log\,(L_\beta^{(k+1)})}{d\beta} = \sum_{(h,h')} d_{hh'}^{(k+1)} X_{hh'} - d\frac{\sum_{(h,h')} X_{hh'} e^{X_{hh'}^T \beta} p_h^{(k)} p_{h'}^{(k)}}{\sum_{(h,h')} e^{X_{hh'}^T \beta} p_h^{(k)} p_{h'}^{(k)}}$$

와 같이 표현할 수 있다. 위의 최대화 방법은 로지스틱 회귀식과 유사성이 있기 때문에 위의 방법은 수치적으로 안정적임을 알 수 있다. 또한 β를 갱신할 때, 뉴튼-랍슨(Newton-Raphon) 알고리즘이나 피셔의 스코어링(Fisher's Scoring) 알고리즘과는 달리 준-뉴튼 알고리즘은 헤시안행렬(Hessian matrix)이나 정보행렬(information matrix)의 역행렬이 필요 없다는 장점도 가지고 있다.

CM-step : update p

위의 CM-step에서 β를 갱신할 때와는 달리 p를 갱신할 때에는 보다 많은 단계를

거치게 되는데, 로그우도는 p에 따라 결정되며 식은 아래와 같다.

$$log\,(L_p^{(k+1)}) \propto \sum_h m_h^{(k+1)} log\,(p_h) - dlog\,(\sum_{(h,h)} e^{X_{hh'}^T \beta^{(k+1)}} p_h p_{h'})$$

위 로그우도의 최대화하기 위한 스코어 식은 $\sum_h p_h = 1$이라는 조건하에

$$\frac{m_h^{(k+1)}}{p_h} - 2d\frac{\sum_h e^{X_{hh'}^T \beta^{(k+1)}} p_{h'}}{\sum_{(h1,h2)} e^{X_{h1h2}^T \beta^{(k+1)}} p_{h_1} p_{h_2}} - p_h = 2n - 2d = 2c$$

이며, 2c, 2d, 2n은 각각 대조군, 환자군, 전체 데이터의 일배체 개수이다.
위 식은

$$u\,(p, \beta^{(k+1)}) = \frac{\sum_h e^{X_{hh'}^T \beta^{(k+10}} p_{h'}}{\sum_{(h,h)} e^{X_{hh'}^T \beta^{(k+1)}} p_h p_{h'}}$$

일 때,

$$p_h = \frac{m_h}{2c + 2du\,(p, \beta^{(k+1)})}$$

로 다시 쓸 수 있다.

여기서 초기값으로 $p^{(k)} \equiv p^{(k,0)}$으로 하고, $p_h^{(k,s+1)} \propto \frac{m_h}{2c + 2du\,(p^{(k,s)}, \beta^{(k+1)})}$을

이용하면 $p_h^{(k,s+1)}$을 구할 수 있다. m_h와 u가 항상 음수가 아니기 때문에 알고리즘을 통해 p를 추정한 값은 항상 적절한 밀도 함수(proper density function)를 가지게 된다.

2.4 EE 알고리즘

2.4.1 EE 알고리즘 개요

최대우도를 이용한 여러 방법들 중에 EM 알고리즘이 가장 많이 쓰이는 방법이기는 하나 SNP 수가 20개 이상인 경우 EM 알고리즘은 일배체를 추정하는데 어려움이 있고, 정보행렬이 비정칙(singular)일 경우 표준오차(standard error)를 추정하지 못하기 때문에 Excoffier는 표준오차를 추정하기 위해서 부스트랩(bootstrap) 방법을 사용하였다(Schneider et al., 2000).

EE(estimating equation) 알고리즘은 다음의 세 가지 면에서 EM 알고리즘을 개선하였는데, 첫째, EE 알고리즘은 SNP 수가 많은 유전형 자료를 다룰 수 있으며, 둘째, 결측값이 존재하는 유전형 자료도 효과적으로 추정할 수 있다. 셋째, 부스트랩 방법을 사용하지 않고 추정된 일배체 빈도에서 표준오차를 계산할 수 있다.

2.4.2 EE 알고리즘 절차

EE 알고리즘은 일배체의 phase가 알려져 있지 않기 때문에 phase 표시자(indicator)를 잠재변수(latent variable)로 생각하는데 n명의 개인 자료가 있을 때, $g_i = (g_{i1}, g_{i2}, \cdots g_{iq})$가 i번째 개인의 SNP이라고 가정하면 $g_{ij} = 0, 1, 2$이다. 또한 p_{ij}를 i번째 개인의 j번째 유전소의 phase 표시자라고 하면 $p_{ij} = 0$ 은 g_{ij}이 하나의 염색체(chromosome)에서 왔음을 의미하는 것이고 $p_{ij} = 1$은 g_{ij}가 다른 부모 염

색체에서 왔음을 의미하는 것이다. $p_i = (p_{i1}, p_{i2}, \cdots, p_{iq})$라 하고 주어진 일배체 빈도 $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_T)$ 하에서 우도함수는 g_i의 분포함수의 곱이 된다. 자료의 개수가 n명일 때, 우도함수를 식으로 표현하면 아래와 같다.

$$L(\theta) = \prod_{i=1}^{n} f(g_i;\theta\ ) \tag{1}$$

여기에 phase정보를 결합하면 p_i와 결합분포(joint distribution)의 주변분포(marginal distribution)로써 g_i의 분포를 다시 쓸 수 있다.

$$f(g_i;\theta) = \sum_{p_i} f(g_i, p_i;\theta) = \sum_{p_i} f(g_i \mid p_i;\theta) f(p_i) \tag{2}$$

위의 식에서 $f(p_i)$는 사전분포(prior distribution)를 의미하는데 일배체의 모호한 성질 때문에 불확실성이 존재하게 되므로 phase 표시자 들이 서로 간에 독립이라고 가정하는 것이 적절하다. 즉, $f(p_i) = \prod_{j=1}^{q} f(p_{ij})$이다.

phase 정보가 주어졌을 때, g_i는 일배체 쌍$(H_i^1 : H_i^2)$로 표현할 수 있고 HWE 가정 하에서 일배체는 무작위로 쌍을 이루게 되므로

$$f(g_i \mid p_i;\theta) = f(H_i^1, H_i^2;\theta) = f(H_i^1;\theta) f(H_i^2;\theta) \tag{3}$$

이다.

$T = 2^q$가 모든 가능한 일배체의 개수라고 할 때, $\{h_1, h_2, \cdots, h_T\}$는 모든 가능한 일배체를 의미하게 되고 $\sum_{t=1}^{T} \theta_t = 1$ 라고 하면, $f(H_i^k;\theta)$는

$$f(H_i^k;\theta) = \prod_{t=1}^{T} \theta_t^{I\{H_i^* = h_t\}} \tag{4}$$

와 같은 다항분포로 나타낼 수 있다.

식 (2)-(4)에 의해 식 (1) 의 우도함수는 아래와 같이 다시 쓸 수 있다.

$$L(\theta) = \prod_{i=1}^{n} 2^{-c_i} \left[\sum_{p_i} \left(\prod_{t=1}^{T} \theta_t^{I(H_i^1 = h_t) + I(H_i^2 = h_t)} \right) \right]$$

위의 식에서 c_i는 i번째 개인의 이형접합(heterozygous) 유전소들(loci)의 개수에서 1을 뺀 수이다. 위의 우도함수를 살펴보면 EM 알고리즘에서의 우도함수와 근본적으로 같음을 알 수 있는데, 양변에 로그를 취한다음 미분을 하면 아래의 식을 얻게 된다.

$$U(\theta) = \sum_{i=1}^{n} U_i(\theta) = \sum_{i=1}^{n} E_{p_i}(F_i \mid g_i;\theta) = \sum_{i=1}^{n} \sum_{p_i} F_i f(p_i \mid g_i;\theta) = 0$$

위 식에서 $F_i = (F_{i1}, F_{i2}, \cdots, F_{iT})'$이고 $F_{ij} = I(H_i^1 = h_j) + I(H_i^2 = h_j) - 2\theta_j$는 j번째 일배체의 관측빈도와 기대빈도의 차이이고

$$f(p_i \mid g_i;\theta) = \frac{f(g_i \mid p_i;\theta) f(p_i)}{\sum_{p_i} f(g_i \mid p_i;\theta) f(p_i)} = \frac{\theta_{H_i^1} \theta_{H_i^2}}{\sum_{p_i;g_i \mid p_i = (H_i^1, H_i^2)} \theta_{H_i^1} \theta_{H_i^2}} \tag{5}$$

는 g_i가 주어졌을 때 p_i의 사후분포(posterior distribution)이다.

식 (5)는 $\theta = \frac{1}{2n}\sum_{i=1}^{n}\sum_{p_i} I_i f(p_i \mid g_i;\theta)$로 다시 쓸 수 있으며 반복적인 단계를 통해 θ의 추정값을 얻을 수 있다.

EE 알고리즘과 EM 알고리즘은 SNP의 개수가 10이상인 경우 일배체의 빈도를 추정하는데 시간이 많이 걸리게 되며, 20개 이상의 SNP이 존재하는 경우 EM 알고리즘은 추정자체가 불가능할 수도 있다. 그러나 EE 알고리즘은 이러한 계산의 한계를 다음과 그림과 같은 방법으로 보완하고 있다.

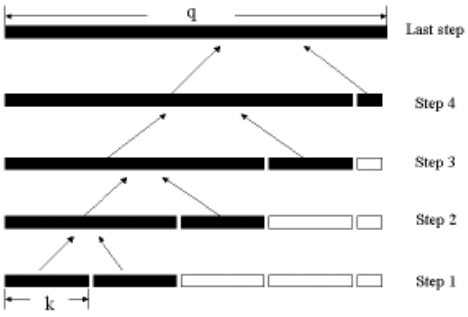

그림 2.1 Forward-block-computational algorithm

즉, SNP의 전체 자료를 k개의 SNP으로 이루어진 블럭(block)으로 나눈 다음, 첫 번째 단계에서는 처음 두개의 블럭에서 일배체 빈도를 추정을 하고 두개의 블럭을 합친(joined) 후 다시 추정을 한다. 각각의 단계에서 빈도수가 극히 작은 일배체는 제거한다. 합쳐진 블럭(joined block)과 그 다음의 하나의 블럭에서 추정하는 과정을 마지막 블럭이 합쳐질 때까지 반복하게 되면 마지막 단계에서의 일배체 빈도 추정값은 전체 SNP 자료에서 빈도를 추정한 값과 비슷하게 된다.

2.5 일배체 빈도 추정 방법 비교

이상의 세 가지 추정방법들은 일배체의 빈도추정을 하는데 있어서 기본적으로

EM 알고리즘을 사용하고 있다. EM 알고리즘은 최대우도를 이용하여 일배체의 확률을 최적화하는 방법인데 여러 모의 실험 결과 HWE를 따르지 않는 자료라 할지라도 성능이 저하되지 않는다는 장점이 있다고 알려져 있다. 반면 일배체를 구성하는 SNP의 수가 많아지거나 유전소의 수가 많아지면 추정이 불가능하다는 단점도 존재한다. 이러한 단점을 극복하기 위해서 ECM, EE 알고리즘과 같은 방법들이 제안되었고 모수 추정단계를 세분화 하여 EM 알고리즘의 단점을 보완한다.

제 3장 선형 모형에 기초한 관련성 분석 방법

3.1 개요

질병과 일배체의 관련성을 보기위한 초기의 접근방법 중의 하나는 카이제곱검정을 이용하는 것이었다. 그러나 카이제곱검정은 분석의 용이성이라는 측면에서의 장점은 있지만 일배체의 수가 많은 경우에는 귀무가설을 기각할 확률이 높아져서 적절치 못한 방법이고 환경적인 요인들을 분석모형에 포함할 수 없다는 단점이 있다. 이러한 단점을 보완하기 위하여 선형모형을 이용한 방법이 제안되었고 구체적인 내용은 다음에서 논의하기로 한다.

3.2 HTR(haplotype trend regression) 방법

3.2.1 HTR 방법 개요

HTR 방법은 혈연관계가 없는 개인에서 선형모형을 이용하여 일배체와 형질간의 관련성을 분석하는 방법인데 EM 알고리즘을 통하여 일배체의 빈도를 추정하고 추정된 일배체의 사후확률과 형질간의 관계를 단순선형모형을 통해 분석한다.

HTR 방법에서는 환자-대조군 자료에서 추정된 일배체를 $2 \times L$ 테이블로 나타낼 수 있다. 두 열은 질병유무를 나타내고 L개의 행은 N개의 개인에서 얻어진 일배체의 수를 의미하며 테이블에서의 각 셀은 EM 알고리즘을 통해 추정된 일배체의 수로 채워지고 한 개인당 두개의 일배체를 가지므로 셀의 빈도를 합하면 총 $2N$개가 된다. 만약 일배체가 모호하지 않다면 환자-대조군자료에서 대립유전자의 빈도 차이를 검정하는 카이제곱검정과 비슷하게 검정할 수 있으며 우도비검정이

나 피어슨의 적합도검정통계량(Pearson's goodness-of-fit statistic)을 이용할 수도 있다.

3.2.2 HTR 모형

$\boldsymbol{Y}$가 형질을 나타내고 $\boldsymbol{D}$가 유전요인을 숫자로 표현한 것이라 하면 HTR방법은

$$\boldsymbol{Y} = \boldsymbol{D}\beta + \epsilon$$

로 표현할 수 있는 N차원의 선형모형이다. 여기서 i번째 개인의 형질은 Y_i이고

$$\boldsymbol{Y^T} = (Y_1, Y_2, ..., Y_N),$$
$$\boldsymbol{D^T} = (\boldsymbol{D_1}, \boldsymbol{D_2}, \cdots, \boldsymbol{D_N}), \boldsymbol{D_i^2} = (D_{i1}, D_{i2}, \cdots, D_{iL})$$

이다. $D_{ij} = 1, \frac{1}{2}, 0$은 각각 i번째 개인이 일배체 j에 대해서 동형접합, 이형접합(heterozygous), 그 밖의 경우일 때를 의미한다. F검정통계량은

$$F = \frac{SSA/(L-1)}{SSE/(N-L)}$$

로 구할 수 있으며 SSA와 SSE는 각각 다음과 같다.

$$SSA = \boldsymbol{Y^T}(\boldsymbol{D}(\boldsymbol{D^T D})^{-1}\boldsymbol{D^T} - \frac{1}{N}\boldsymbol{J_{N\times N}})\boldsymbol{Y}, SSE = \boldsymbol{Y^T}(\boldsymbol{I_N} - \boldsymbol{D}(\boldsymbol{D^T D})^{-1}\boldsymbol{D^T})\boldsymbol{Y}$$

HTR 방법을 설명하기 위하여 간단한 예를 생각해 보자. 3명의 개인이 각각 형질 Y_1, Y_2, Y_3을 가지고 있고 각각의 일배체는 $h_1/h_1, h_2/h_3, h_1/h_3$이며 모호하지

않다고 가정하면 $E(\boldsymbol{Y}) = \boldsymbol{D}\beta$는 아래와 같이 나타낼 수 있다.

$$E(Y) = E\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{array}{c} \begin{matrix} \mu & h_1 & h_2 & h_3 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1/2 & 1/2 \\ 1 & 1/2 & 0 & 1/2 \end{bmatrix} \end{array} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

그러나 일배체가 모호하다면 위의 행렬 D는 유전형이 주어졌을 때 일배체의 조건부확률이 된다. 예를 들어 일배체 h_2와 h_3의 경우를 생각해 보면, 유전형이 G_i일 때 일배체 쌍(h_2, h_3)이 나타날 조건부확률은

$$P(h_2, h_3 | G_i) = \frac{P(G_i | h_2, h_3) p_{h_2} p_{h_3}}{\sum_{u,v} P(G_i | h_u, h_v) p_{h_u} p_{h_v}}$$

이며, p_{h_u}와 p_{h_v}는 일배체의 빈도를 나타낸다. 따라서 모호한 일배체에서는 행렬 D는 더 이상 0, 1/2, 1의 값을 가지지 않는 것이다.

3.3 스코어 방법

3.3.1 스코어 방법 개요

스코어 방법은 일배체의 빈도를 추정하기 위해 EM 알고리즘을 사용하였고 일반화선형모형(generalized linear models ; GLMs)을 사용하여 다양한 형태의 형질에 대해서 일배체의 관련성 분석이 가능하도록 하였으며, 환경요인이 형질에 미치는 영향을 고려한 통계량을 구할 수 있다.

3.3.2 일반화선형모형

일반화선형모형을 사용하기 위해 유전형이 모호하지 않다고 가정해 보자. y를 관측된 형질, $\boldsymbol{X_e}$를 관측된 환경요인들의 벡터, $\boldsymbol{X_g}$를 유전요인을 숫자로 표현한 벡터라고 한다. 공변량이 형질의 평균에 영향을 미치고 척도(sclae)에는 영향을 미치지 않는다고 가정하면 공변량들의 효과들은, $\boldsymbol{\alpha}$를 절편과 환경요인의 회귀계수라 하고 $\boldsymbol{\beta}$를 유전요인의 효과라고 할때,

$$\eta = \boldsymbol{X_e'\alpha} + \boldsymbol{X_g'\beta}$$

와 같은 선형함수로 표현할 수 있고 유전요인과 형질간의 관련성이 없다는 것은 $H_0 : \boldsymbol{\beta} = 0$을 검정함으로서 알 수 있다.

$\boldsymbol{Z} = (\boldsymbol{X_e} \mid \boldsymbol{X_g})$, $\boldsymbol{\gamma} = (\boldsymbol{\alpha} \mid \boldsymbol{\beta})$라고 하면 $\boldsymbol{Z}$가 주어졌을 때 y의 우도는 지수족(exponential family data)에 대한 일반화 선형모형으로 표현할 수 있다. 식은 아래와 같다.

$$L(y \mid \boldsymbol{Z}) = \exp\left[\frac{y\eta - b(\eta)}{a(\phi)} + c(y, \phi)\right]$$

위 식에서 a, b, c는 분포에 따라 결정되고 ϕ는 산포모수(dispersion parameter), $\eta = \boldsymbol{Z\gamma}$ 인 정준연결함수(canonical link function)이다. 분포에 따른 일반화선형모형의 함수들은 다음 표와 같다.

표 3.1 지수족(exponential family)에서의 일반화선형모형의 분포들의 특징

	정규분포	포아송분포	이항분포	감마분포	역가우스분포
정의	$N(\mu,\sigma^2)$	$P(\mu)$	$B(m,\pi)/m$	$G(\mu,\nu)$	$IG(\mu,\sigma^2)$
y의 범위	$(-\infty,\infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0,\infty)$	$(0,\infty)$
산포모수 ϕ	σ^2	1	$1/m$	ν^{-1}	σ^2
누적함수 $b(\eta)$	$\theta^2/2$	$exp(\theta)$	$log(1+e^{\theta})$	$-log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y,\theta)$	$-\frac{1}{2}(\frac{y^2}{\phi}+log(2\pi\phi))$	$-logy!$	$\log\binom{m}{my}$	$\nu log(\nu y)-logy-log\Gamma(\nu)$	$-\frac{1}{2}\{log(2\pi\phi y^3)+\frac{1}{\phi y}\}$
$\mu(\theta):E(Y;\theta)$	θ	$exp(\theta)$	$e^{\theta}/(1+e^{\theta})$	$-1/\theta$	$(-2\theta)^{-1/2}$
정준연결 $\theta(\mu)$	identity	log	logit	reciprocal	$1/\mu^2$
분산함수 $V(\mu)$	1	μ	$\mu(1-\mu)$	μ^2	μ^3

3.3.3 스코어 모형

유전형이 모호하지 않다고 가정하면 벡터 $\boldsymbol{Z}$에 대한 스코어 통계량은

$$\boldsymbol{U_\gamma} = \sum_{i=1}^{N} \frac{\partial ln(L_i)}{\partial \gamma} = \sum_{i=1}^{N} \frac{y_i - \widetilde{y_i}}{a(\phi)} \boldsymbol{Z_i}$$

이고 $\widetilde{y_i}$는 공변량 벡터 $\boldsymbol{Z_i}$에 의해 적합된 값(fitted value)이며 $\boldsymbol{\gamma}$는 회귀계수, $\boldsymbol{N}$은 자료의 개수이다. 스코어 방법은 환경요인을 제어할 수 있는데 먼저 유전요인에 대한 효과 $\boldsymbol{\beta}$를 0으로 하고 $\hat{\alpha}$를 추정한 다음 $\boldsymbol{U_{\hat{\alpha}}} = 0$이 되도록 $\widetilde{y_i}$를 결정하면 환경요인을 제어한 상태에서의 유전요인과 형질간의 관계를 볼 수 있다.
이 때 스코어 통계량은

$$\boldsymbol{U_\beta} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \frac{\boldsymbol{y_i - \widetilde{y_i}}}{\boldsymbol{a(\phi)}} \boldsymbol{X_{g_i}}$$

이고 귀무가설 하에서 $\boldsymbol{U_\beta}$의 분산은

$$\begin{aligned} \boldsymbol{V_\beta} &= \boldsymbol{V_{\beta\beta} - V_{\beta\alpha} V_{\alpha\alpha^{-1}} V_{\alpha\beta}} \\ &= \left[\frac{b'(\eta)}{a(\phi)}\right] \left(\sum_{i=1}^{N} X_{gi} X_{gi}' - \frac{X_g . X_g .'}{N}\right) \end{aligned}$$

이다.

모호한 일배체에 있어서도 위와 같은 방법으로 생각하면 환경요인을 제어한 상태에서의 유전요인과 형질간의 관계를 볼 수 있는데, 스코어 통계량은

$$U_\beta = \sum_{i=1}^{N} \frac{y_i - \widetilde{y_i}}{a(\phi)} E_p(X_{g_i})$$

이다. $E_p()$는 표식유전자 정보가 주어졌을 때, 귀무가설 하에서의 유전형의 사후 분포에 대한 기대값을 의미하며

$$E_p(X) = \sum_{g \in G} X_g Q(g)$$

로 구할 수 있다. 유전형에 대한 사후확률은 $Q(g) = P(g) / \sum_{g \in G} P(g)$이며, 유전형에 대한 확률 $P(g)$는 EM 알고리즘을 이용하여 추정한다. U_β의 분산은 $V_\beta = V_{\beta\beta} - V_{\beta\alpha} V_{\alpha\alpha}^{-1} V_{\alpha\beta}$로 나타낼 수 있으며, $V_{\alpha\alpha}$, $V_{\alpha\beta}$, $V_{\beta\beta}$는 각각 다음과 같이 구할 수 있다.

$$\begin{aligned}
V_{\alpha\alpha} &= \sum_{i=1}^{N} \frac{b''(\eta_i)}{a(\phi)} X_{ei} X_{ei}', \\
V_{\alpha\beta} &= \sum_{i=1}^{N} \frac{b''(\eta_i)}{a(\phi)} X_{ei} E_p(X_{ei}'), \\
V_{\beta\beta} &= \sum_{i=1}^{N} \left[\frac{b''(\eta_i)}{a(\phi)} - \frac{(y_i - \widetilde{y_i})^2}{a(\phi)^2} \right] E_p(X_{gi} X_{gi}') \\
&\quad + \frac{(y_i - \widetilde{y}_i)^2}{a(\phi)^2} E_p(X_{gi}) E_p(X_{gi}')
\end{aligned}$$

스코어 방법은 전체 일배체를 고려한 경우와 개별적인 일배체를 고려한 경우, 각각의 스코어 통계량을 구할 수 있다는 장점이 있는데 전체적 스코어 통계량 (global score statistic)은

$$S = U_\beta' V_\beta^{-1} U_\beta$$

이며 대표본일 때 자유도가 V_β의 계수(rank)인 카이제곱분포를 따른다. 또한 스코어 통계량은 우도비검정통계량(likelihood-ratio test statistic)과 근사적으로 같으나 β에 대한 최대우도추정값을 구하지 않아도 되므로 계산시간이 빠르다는 장점도 가지고 있다. 개별 일배체에 대한 스코어 통계량은 X_g의 k번째 성분(component)을 계산함으로서 얻을 수 있는데 $z_k = U_{\beta,k} / \sqrt{V_{\beta,k,k}}$로 표현할 수 있으며 자유도가 1인 카이제곱분포를 따르고 대표본에서는 근사적으로 표준정규분포(standard normal distribution)를 따른다.

만일 특정한 k번째 일배체의 효과가 다른 모든 일배체효과보다 크다면 z_k^2는 최대화될 것이고 전체적 스코어 통계량을 사용했을 경우보다 검정력이 크게 될 것이다. 그러나 불완전한 LD(linkage disequillibrium)때문에 하나의 특정 일배체보다는 몇 개의 일배체가 동시에 작용하여 형질과 관련성을 보이는 상황이 유전학적으로 보다 일반적인 때문에 전체적 스코어 통계량을 사용하는 것이 바람직한 경우가 많다.

3.4 Chaplin 방법

3.4.1 Chaplin 방법 개요

Chaplin 방법은 다른 방법과는 달리 질병에 관련이 있을 것으로 보이는 특정 일배체의 우성(dominant), 열성(recessive), 공동우성(codominant)의 효과를 볼 수 있고 환자군이 HWE 상태가 아니어도 된다. 또한 EM 알고리즘을 응용한 ECM 알고리즘을 사용하여 결측치를 다룰 수 있다. Chaplin 방법은 ECM 알고리즘을 이용하여 일배체의 빈도를 추정하고 로지스틱회귀모형을 이용하여 질병과 일배체

의 관련성을 분석한다.

3.4.2 Chaplin 방법에서의 일배체 빈도추정

혈연관계가 없는 n명의 개인으로 이루어진 표본이 대조군 c명, 환자군 d명으로 이루어져 있다고 하자. D가 질병유무를 나타낸다고 하면 $D=0,1$의 값을 가지고 $G=g$는 개인의 다중 SNP(multi-SNP)를, $H=(h,h')$는 개인의 일배체 쌍을 의미한다고 하자. 여기서 개인이 2개 이상의 SNP에서 이형접합이거나 한개 이상의 SNP가 결측되면 일배체 쌍 $H=(h,h')$는 알 수 없다. $S(g)$를 일배체 쌍 $\{H=(h,h')\}$들의 집합이라고 하면 $h \neq h'$에 대해서 $(h,h') \in S(g)$는 $(h',h) \in S(g)$와 같은 의미를 갖게 된다.

Chaplin 방법은 자료의 우도를 이용하는 방법인데, 관측된 유전형 자료의 우도를 질병유무가 주어졌을 때의 일배체의 함수로 나타내고 식은 다음과 같다.

$$L_{OBS} = \prod_{g} [P(G=g|D=0)]^{c_g} \, [P(G=g|D=1)]^{d_g}$$

$P(G=g|D=0)$과 $P(G=g|D=1)$은 각각 대조군과 환자군에서 유전형이 g일 확률이고 c_g와 d_g는 자료에서 유전형 g를 가지고 있는 대조군과 환자군의 수이다.

$P(G=g|D=0)$과 $P(G=g|D=1)$를 일배체 쌍의 빈도라고 하면 L_{OBS}를 일배체 쌍의 함수로도 나타낼 수 있다. $\pi_{hh'} = P[H=(h,h')|D=0]$, $\rho_{hh'} = P[H=(h,h')|D=1]$이라고 하면 $\pi_{hh'}$와 $\rho_{hh'}$는 각각 대조군과 환자군에서의 일배체 쌍 $H=(h,h')$의 빈도를 의미하게 되며

$$P[G=g|D=0] = \sum_{(h,h') \in S(g)} \pi_{hh'}$$

$$P[G=g|D=1] = \sum_{(h,h') \in S(g)} \rho_{hh'}$$

이므로

$$L_{OBS} = \prod_{g} \left(\sum_{(h,h') \in S(g)} \pi_{hh'} \right)^{c_g} \left(\sum_{(h,h') \in S(g)} \rho_{hh'} \right)^{d_g}$$

으로 우도함수를 다시 쓸 수 있다. 특정 일배체의 특성의 추정을 용이하게 하기 위해

$$\theta_{hh'} = \frac{P[D=1|H=(h,h')]}{P[D=0|H=(h,h')]}$$

라고 정의하면

$$\rho_{hh'} = P[H=(h,h')|D=1]$$

$$= \frac{P[H=(h,h'),D=1]}{\sum_{(h_1,h_2)} P[H=(h_1,h_2),D=1]}$$

$$= \frac{\theta_{hh'} P[H=(h,h'),D=0]}{\sum_{(h_1,h_2)} \theta_{h_1h_2} P[H=(h_1,h_2),D=0]}$$

$$= \frac{\theta_{hh'}\pi_{hh'}}{\sum_{(h_1,h_2)} \theta_{h_1h_2}\pi_{h_1h_2}}$$

이다. 결과적으로 $\pi_{hh'}$와 $\theta_{hh'}$를 정함으로서 $\rho_{hh'}$를 결정할 수 있게 된다. 위의 식을 이용하면 우도함수 L_{OBS}를 다시 쓸 수 있으며 식은 다음과 같다.

$$L_{OBS} = \frac{\prod_g (\sum_{(h,h') \in S(g)} \pi_{hh'})^{c_g} (\sum_{(h,h') \in S(g)} \theta_{hh'} \pi_{hh'})^{d_g}}{(\sum_{(h_1,h_2)} \theta_{h_1 h_2} \pi_{h_1 h_2})^d}$$

그러나 위에서 구한 L_{OBS}를 이용하는 것은 일배체가 모호할 경우에 문제가 생길 수 있는데 주어진 SNP으로부터 일배체를 추정할 때 모집단에서 나타날 수 있는 일배체의 수가 몇 가지로 제한되어 있다거나 하는 추가 정보가 없을 경우에는 $S(g)$의 경우의 수가 많아져서 결국 적절한 일배체를 추정할 수 없게 되고 결과적으로 $\pi_{hh'}$와 $\theta_{hh'}$도 추정할 수 없게 된다.

위의 추정상의 문제를 해결하기 위하여 대조군에서 일배체 쌍이 HWE를 만족한다고 가정하면 p_h가 대조군에서의 일배체 h의 빈도라고 할 때, $\pi_{hh'} = P[H = (h,h') | D = 0] = p_h p_{h'}$이다. 또한 β가 질병과의 관련성을 나타내는 모수 벡터이고 R-차원을 가지며, $X_{hh'}$는 설계벡터(design vector)라고 하면 $\theta_{hh'} = e^{X_{hh'}^T \beta}$이며 위의 가정에 의해

$$L_{OBS} = \frac{\prod_g (\sum_{(h,h') \in S(g)} p_h p_{h'})^{c_g} (\sum_{(h,h') \in S(g)} e^{X_{hh'}^T \beta} p_h p_{h'})^{d_g}}{(\sum_{(h,h)} e^{X_{hh'}^T \beta} p_h p_{h'})^d} \quad (6)$$

이고 모호한 반수체의 추정을 위해서는 위의 식을 사용하게 된다.

3.4.3 Chaplin 모형

식 (6)의 우도함수를 이용하여 질병과의 관련성을 나타내는 모수 β의 효과를 검정할 수 있는데 귀무가설 $H_0 : \beta = 0$은 일배체가 질병과 관련성이 없다는 것을 의미한다. 귀무가설을 검정하기 위하여 점근이론(asymptotic theory)에 근거하여 두 가지 검정방법을 고려할 수 있는데, 하나는 우도비 통계량(likelihood-ratio statistic)이고 다른 하나는 로버스트 스코어 통계량(robust score statistic)이다.

우도비 통계량은 $LR = 2log(L_{H_a}/L_{H_0})$으로 표현할 수 있는데 L_{H_0}과 L_{H_A}는 각각 귀무가설과 대립가설 하에서의 L_{OBS}값이다. 귀무가설 하에서 우도비 통계량은 근사적으로 카이제곱분포를 따르며 자유도는 검정에 사용된 회귀계수의 수와 같다.

β의 효과를 검정하기 위해서 사용될 수 있는 다른 방법은 로버스트 스코어 통계량을 사용하는 것이다. Chaplin 방법은 Boos(1992)가 제안한 통계량을 사용하였으며 β와 p_h에 대하여 각각 로그우도를 미분하는 과정이 필요하다. β에 대한 로버스트 스코어 통계량은

$$\frac{\partial log(L_{OBS})}{\partial \beta} \equiv U_\beta = \sum_g d_g(\overline{X}_g - \overline{X})$$

이고 $\overline{X}_g$와 $\overline{X}$는 각각 다음과 같다.

$$\overline{X}_g = \frac{\sum_{(h,h') \in S(g)} p_h p_{h'} e^{X_{hh'}^T \beta} X_{hh'}}{\sum_{(h,h') \in S(g)} p_h p_{h'} e^{X_{hh'}^T \beta}},$$

$$\overline{X} = \frac{\sum_{(h,h')} p_h p_{h'} e^{X_{hh'}^T \beta} X_{hh'}}{\sum_{(h,h')} p_h p_{h'} e^{X_{hh'}^T \beta}}$$

와 같다.

3.5 H-plus 방법

3.5.1 H-plus 방법 개요

H-plus 방법은 스코어 방법과 마찬가지로 유전 요인과 환경 요인을 모델에 포함할 수 있는데 일배체가 모호하다는 가정 하에 EE 알고리즘을 이용하여 일배체의 분포를 추정한 다음 질병과 두 요인이 어떤 관련성을 보이는지를 로지스틱회귀분석을 통해 분석한다. Chaplin 방법과 마찬가지로 특정 일배체의 효과를 우성, 열성, 공동우성의 세 가지 측면에서 분석할 수 있으며 일배체와 다른 환경 요인들을 공변량으로 봤을 때, 질병과의 관련성을 침투함수(penetrance function)로 표현한 다음 로지스틱회귀식을 이용하여 형질과 일배체, 환경요인을 모델화한다. 구체적인 방법은 다음과 같다.

3.5.2 H-plus 모형

$I(h_i^1, h_i^2, x_i, \beta)$가 일배체 (h_i^1, h_i^2), 공변량 (x_i), 계수 β의 함수라고 하면 로지스틱 침투함수는 다음과 같다.

$$P(d_i = 1 | h_i^1, h_i^2, x_i) = \frac{1}{1 + exp\,[-\alpha - I(h_i^1, h_i^2, x_i, \beta)]} \qquad (7)$$

위 함수는 0에서 1사이의 값을 갖고, 질병에 걸렸을 확률을 양적으로 표현하고 함수 $I(h_i^1, h_i^2, x_i, \beta)$는 귀무가설에 따라 선택한다. $I(\cdot)$을 선택하는 구체적인 방법은 다음과 같다.

첫째, 일배체의 효과를 보고자 할 때

$$I(h_i^1, h_i^2, x_i, \beta) = \beta_1{}'[K(h_i^1) + K(h_i^2)] + \beta_2{}' x_i$$

라고 할 수 있다. 여기서 $K(\cdot)$는 일배체를 벡터로 표현한 것이고 l번째 일배체가 질병과 관련이 없음을 보려면 귀무가설 $\beta_{1l} = 0$를 검정하면 된다.

둘째, 이배체의 효과를 보고자 할 경우이다. 일반적인 경우 일배체의 효과가 분석하고자 하는 관심대상인 경우가 많으나 질병과의 관련성이 일배체가 아니라 일배체쌍인 이배체와 있을 수 있다는 것이다. 이러한 이배체와의 관련성은 침투모드(penetrance mode)에 따라서 우성(dominant), 열성(recessive), 공동우성(codominant)의 3가지로 분류된다. h가 분석하고자 하는 일배체라고 가정하면 3가지 침투모드에 따라 다음과 같이 $I(h_i^1, h_i^2, x_i, \beta)$의 선택이 달라진다.

dominant mode :

$$I(h_i^1, h_i^2, x_i, \beta) = \beta_1{}' K(h_i^1, h_i^2) + \beta_2{}' x_i$$
$$K(h_i^1, h_i^2) = \begin{cases} 1 & h_i^1 = \tilde{h} \ or \ h_i^2 = \tilde{h} \\ 0 & otherwise \end{cases}$$

recessive mode :

$$I(h_i^1, h_i^2, x_i, \beta) = \beta_1' K(h_i^1, h_i^2) + \beta_2' x_i$$
$$K(h_i^1, h_i^2) = \begin{cases} 1 & h_i^1 = \tilde{h} \text{ and } h_i^2 = \tilde{h} \\ 0 & otherwise \end{cases}$$

codominant mode :

$$I(h_i^1, h_i^2, x_i, \beta) = \beta_{11}' K(h_i^1, h_i^2) + \beta_{12}' K(h_i^1, h_i^2) + \beta_2' x_i$$
$$K(h_i^1, h_i^2) = \begin{cases} 1 & h_i^1 = \widetilde{h_j} \text{ and } h_i^2 = \widetilde{h_j} \\ 0 & otherwise \end{cases}, j = 1, 2$$

셋째, 일배체와 공변량의 교호작용을 보고자 할 때는 $\beta_3 = (\beta_{31}, \beta_{32}, ...)'$ 이라고 할 때,

$$I(h_i^1, h_i^2, x_i, \beta) = \beta_1' [K(h_i^1) + K(h_i^2)] + \beta_2' x_i + \beta_3' [K(h_i^1) + K(h_i^2)] x_i$$

처럼 모델을 세울 수 있고 모든 후보 일배체와 하나의 공변량간의 교호작용을 수치로 표현할 수 있다.

식 (7) 에서 계수 β는 환자-대조군의 자료로부터 추정되는 것이므로 질병에 걸릴 확률은 다음과 같이 표현할 수 있다.

$$\mu_i = f(d_i = 1 | h_i^1, h_i^2, x_i) = \frac{1}{1 + exp[-\xi - I(h_i^1, h_i^2, x_i, \beta)]}$$

여기서 절편 $\xi = \alpha + log\frac{(1-\theta)\eta}{\theta(1-\eta)}$는 로짓척도(logit scale)에 의해 이동된 절편이고 θ는 환자군의 비율, η는 전체 모집단에서의 질병에 걸릴 확률을 나타낸다

(Prentice and Pyke, 1979). 계수를 추정할 때 주의할 점은 각 일배체마다 하나의 계수가 존재하므로 빈도가 높은 일배체(common haplotype)의 수가 많을수록 추정해야 할 계수의 수도 많아진다는 점이다.

3.6 선형모형에 기초한 질적형질과 일배체의 분석 방법 비교

앞에서 논의한 네 가지 방법의 모형과 통계량을 요약하여 비교하면 다음 표와 같다.

표 3.2 일배체를 이용한 분석방법들의 모형과 통계량 비교

Method	Linear Model	Statistic
HTR	Simple Regression	$F = \frac{SSA/(L-1)}{SSE/(N-L)}$
Score	Generalized Linear Model	$S = U_\beta' V_\beta^{-1} U_\beta$
Chaplin	Logistic Regression	Likelihood Statistic $LR = 2log(L_{H_a}/L_{H_0})$ Robust Score Statistic $U_\beta = \sum_g d_g(\overline{X}_g - \overline{X})$
H-plus	Logistic Regression	$LR = 2log(L_{H_a}/L_{H_0})$

환자-대조군 자료의 경우 설계(design)측면에서는 표본추출이 질병 유무와 같은 기준으로 이루어진다는 면에서 후향적으로 확인된(retrospectively ascertained) 자료이나 본 논문에서는 분석방법의 차이를 기준으로 전향적(prospective), 후향적(retrospective)모델로 나누어 보기로 한다.

HTR, 스코어, H-plus 방법들은 질병의 유무나 질병과 관련이 있는 양적형질이 반응변수(response variable)로, 일배체는 설명변수(explanatory variable)로 다루어진다는 면에서 전향적 모델이라고 볼 수 있으며 Chaplin 방법은 질병 유무가 주어졌을 때 일배체의 분포를 결과로 보고 조건부 확률을 구한다는 점에서 후향적 모델이라고 생각할 수 있다. 네 가지 방법을 간략히 비교하면 표 3.3과 같다.

표 3.3 일배체를 이용한 분석방법 비교

Method	Trait	Likelihood	HWE	Covariates	Global test	Reference
HTR	Binary/ quantitative	prospective	Pool	No	Yes	Zaykin et al., 2002
Score	Binary/ quantitative	prospective	Pool	Yes	Yes	Schaid et al., 2002
Chaplin	Binary	retrospective	Controls	No	Yes	Epstein and Satten et al., 2003
H-plus	Binary	prospective	Controls	Yes	No	Zaho et al., 2003

Prentice and Pyke(1979)는 후향적 자료라 할지라도 전향적 방법으로 분석하는 것이 유효하다는 것과 전향적 분석 방법이 후향적 분석 방법과 마찬가지로 효율적임을 보였다. 그러나 이러한 결과는 종속변수가 포화된 비모수적 분포(saturated non-parametric distribution)을 따른다는 가정 하에서만 얻을 수 있고, 만약 종속변수의 분포에 어떠한 제약 사항이 가해질 경우에는 전향적 분석 방법은 후향적

분석 방법보다 효과적이지 않다(Carroll et al., 1995). Carroll의 분석 결과를 일배체 분석에 적용해 보면 HWE 가정사항이 제약 사항이 될 수 있을 것이다. 대부분의 일배체 분석 방법들은 어떠한 형태로든지-전체 자료에서 혹은 환자군 자료에서 - HWE를 가정하므로 본질적으로 전향적인 분석 방법은 후향적 분석 방법에 비해 효율적이지 못할 수도 있다. 그러나 후향적인 방법은 환경요인을 모델에 포함시킬 수 없다는 점과 HWE 가정위배 유무에 결과가 달라진다는 점이 단점으로 작용한다. 그에 반해서 전향적인 방법-특히 스코어 방법과 H-plus 방법-들은 HWE 가정에 덜 민감하다(robust)는 장점이 있다. 다음 장에서는 제 3장에서 논의한 내용들을 실재자료를 이용하여 비교해 보기로 한다.

제 4장 실재자료를 이용한 관련성 분석

4.1 개요

실재자료로 일 병원 심혈관계질환 유전체연구센터에서 수집된 유전체자료(표 4.1)를 이용하여 앞 장에서 논의한 네 가지 방법으로 일배체와 고혈압간의 관련성을 분석하였다. 유전체자료는 혈연관계가 없는 개인들을 표본으로 하였고 두개 이상의 SNP이 존재하여 일배체를 구성할 수 있는 10개의 유전체자료를 이용하였다. 전체표본은 3344명이나 개인마다 판독된 유전체자료의 수가 다르므로 일배체마다 분석되는 표본수가 다르며 결측치가 존재하지 않는 완전자료와 결측치를 1개만 허용한 불완전자료의 환자군과 대조군의 수는 표4.2와 같다. 실재분석에 있어서 불완전자료의 경우 SNP수가 3개 이상인 자료만 이용하였는데 SNP수가 2개일 때 결측치가 존재하면 하나의 SNP만으로 일배체를 구성한다는 것은 의미가 없기 때문이다.

실재 분석에 앞서 각 유전체마다 SNP이 HWE 상태인지 확인해 보았다. 완전자료의 경우, 유의수준 0.05에서 APM의 G276T(p-value=0.046)와 APOA1의 G-75A가 귀무가설을 기각하였다(p-value<0.0001). 불완전자료의 경우는 유의수준 0.05에서 APOA1의 G-75A가 귀무가설을 기각하였다.

다음으로 SNP들이 연관불균형(linkage disequilibrium)관계에 있는지 검정해 보았다. 연관균형관계에 놓여 있는 유전체자료는 SNP간의 연관성이 없기 때문에 구성된 일배체의 의미가 없기 때문이다. 귀무가설을 SNP들이 연관균형관계에 놓여 있다고 했을 때, 결측치를 0개 허용했을 경우에는 LIPC를 구성하는 SNP들이 귀무가설을 기각하지 못했고(p-value=0.984), 결측치를 1개 허용했을 경우에도 LIPC는 귀무가설을 기각하지 못했다(p-value=0.982). 따라서 LIPC는 분석에서 제외하고 나머지 9개의 유전체자료만을 실재분석에 사용하였다.

표 4.1 심혈관계질환 유전체연구센터 유전체자료 현황

Gene name		SNP수	Description
Angiotensin I converting enzyme	ACE	6	A2400T(A-240T), C2547T(C-93T), 14094(ALU I/D), G14480C, A14519G, A22982G,
Angiotensinogen	AGT	4	G-217T, A-20C, G-6A, M235T(T/C)
Arachidonate 5-lipoxygenase-activating protein	ALOX5AP	3	T-1414G, IVS1+18C/A, IVS2+105C/T
Adiponectin	APM	2	T45G, G276T
Apolipoprotein A1	APOA1	2	XMNI, G-75A
Apolipoprotein A5	APOA5	5	C-1390T, T-1131C, G-1020A, G-3A, 1259T/C
Cholesteryl ester transfer protein	CETP	3	C-629A,TAQ1B(G/A), I405V(A/G)
Endothelial adhesion molecule 1	ESEL	2	S128R(A/C),H468Y(C/T), L575F(C/T)
Hepatic lipase	LIPC	2	C-514T, V95M(G/A)
Microsomal triglyceride transfer protein	MTP	2	5A/6A,C-1562T

표 4.2 유전체자료의 표본수 비교

	완전자료			불완전자료			표본수 차이
	표본수	환자군	대조군	표본수	환자군	대조군	
ACE	1271	786	485	1332	823	509	61
AGT	1350	839	511	1382	867	515	32
ALOX5AP	115	35	80	121	35	86	6
APM	119	34	85	121	35	86	2
APOA1	631	186	445	1266	795	471	635
APOA5	82	23	59	86	25	61	4
CETP	1169	757	412	1373	857	516	204
ESEL	449	197	252	568	227	281	119
LIPC	659	190	469	1307	833	474	648
MTP	298	50	248	1101	625	476	803

4.2 유전요인을 고려한 관련성 분석 결과

환경요인을 고려하지 않았을 때 구성된 일배체와 질병과의 관련성이 존재하지 않는다는 귀무가설 하에서 전체적인 검정(global test)를 수행하였다. 분석은 완전자료와 불완전자료에서 각각 수행하였다.

4.2.1 완전자료에서 유전요인을 고려한 관련성 분석 결과

완전자료에서 고혈압과 일배체의 관련성 분석을 한 결과를 유의확률(p-value)를 중심으로 비교하였다. 고혈압과 일배체의 관련성이 없다는 귀무가설을 기각하기 위한 유의수준을 0.05라 했을 때, ACE가 관련성이 있는 것으로 나왔다. SNP 수에 따른 특징을 알아보기 위해 표 4.3의 결과를 SNP수에 따라 나누어서 비교해

보았다(그림 4.1 - 그림 4.4). 효과적인 비교를 위하여 유의확률에 $-log_{10}$을 취하였고 $-log_{10}(0.05) = 1.301$이므로 $-log_{10}$를 취한 값이 1.301보다 크면 고혈압이 일배체와 관련성이 있다고 말할 수 있다.

표 4.3 완전자료에서 유전요인을 고려한 관련성 분석 결과

	SNP수	HWE pool	p-value		HWE controls	p-value	
			HTR	Score		Chaplin	H-plus
ACE	6	Yes	0.19	0.015	No	0.029	0.043
AGT	4	Yes	0.62	0.470	Yes	0.486	0.465
ALOX5AP	3	Yes	0.84	0.841	No	0.891	0.886
APM	2	No	0.70	0.709	Yes	0.878	0.636
APOA1	2	No	0.39	0.488	No	0.583	0.456
APOA5	5	Yes	0.07	0.068	Yes	0.141	0.247
CETP	3	Yes	0.21	0.210	No	0.287	0.214
ESEL	2	Yes	0.87	0.930	Yes	0.932	0.826
MTP	2	Yes	0.98	0.812	Yes	0.935	0.928

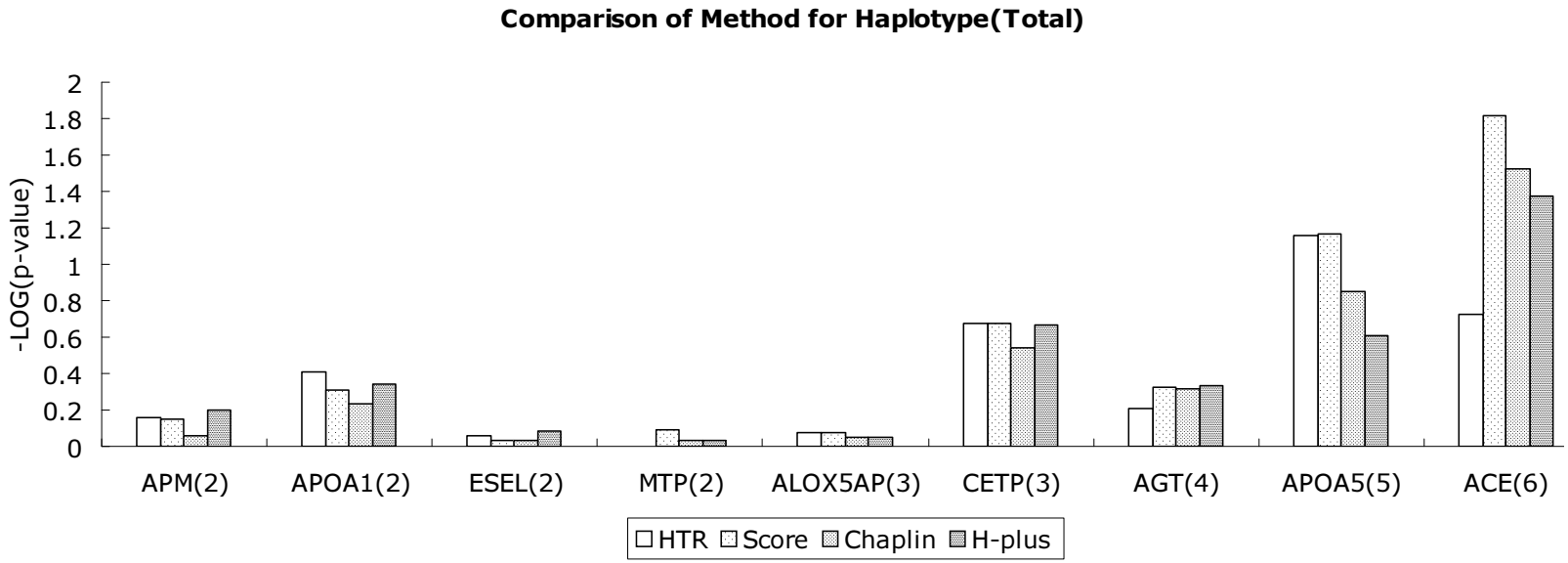

그림 4.1 완전자료에서 고혈압과 일배체의 관련성 분석 결과

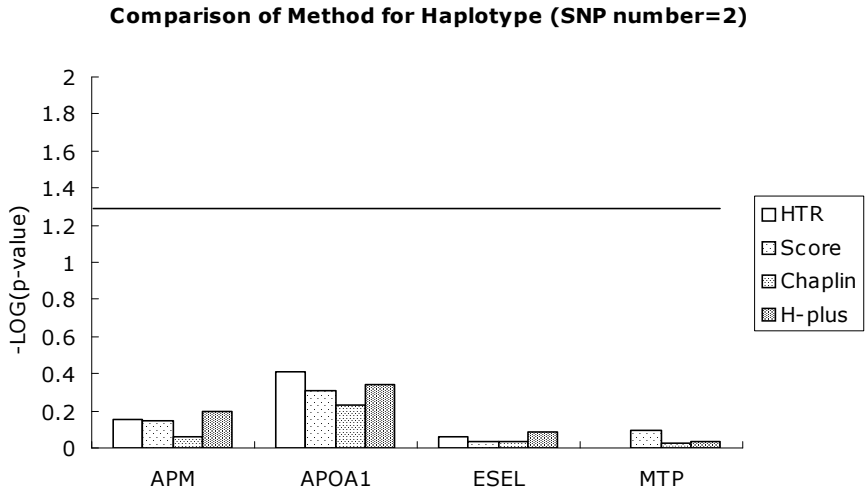

그림 4.2 완전자료에서 2개의 SNP을 이용한 관련성 분석 결과

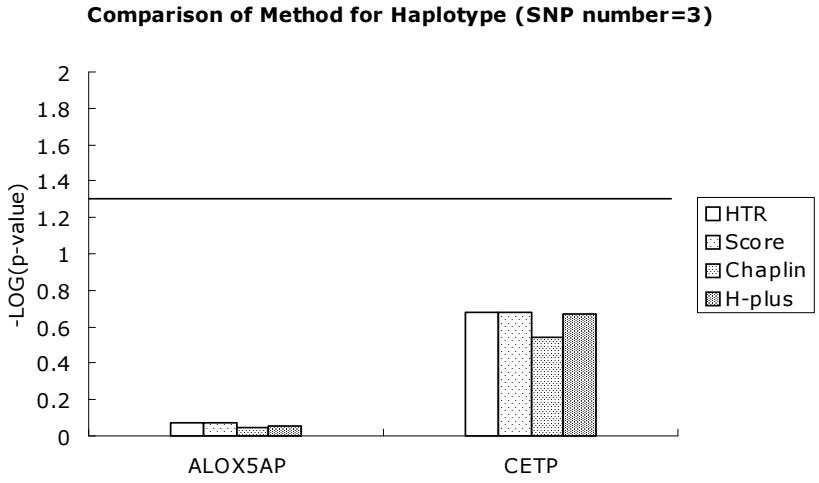

그림 4.3 완전자료에서 3개의 SNP을 이용한 관련성 분석 결과

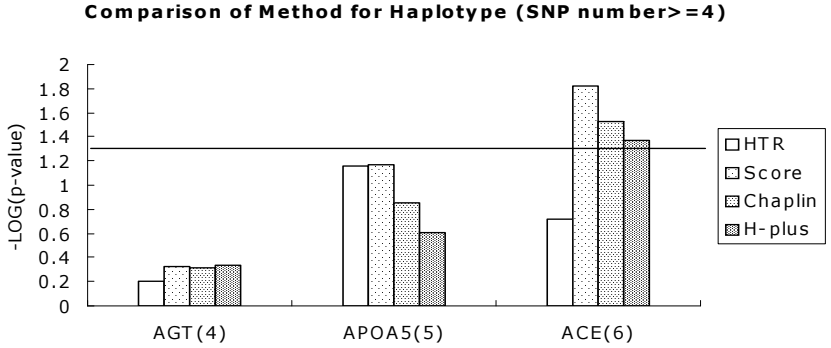

그림 4.4 완전자료에서 4개 이상의 SNP을 이용한 관련성 분석 결과

그림 4.1 - 그림 4.4에서 알 수 있듯이 SNP수가 3개일 때 모든 방법이 비슷한 결과를 보여주었다. SNP수가 2이거나 4이상인 경우에는 방법간의 차이가 있는 것으로 보이는데, HTR방법이 가장 일관되지 못한 결과를 보였고 SNP수가 많아질수록 다른 방법과의 차이가 심해지는 것으로 보인다. 특히 ACE 같은 경우, 다른 모든 방법은 유의한 결과를 보였으나 HTR만이 관련성이 없다는 반대의 결과를 보여주었다. HTR을 제외한 나머지 세 방법만 비교했을 경우는 스코어 방법이 대체적으로 가장 높은 값을 보였다. SNP수와 관계없이 Chaplin 방법은 대체적으로 가장 낮은 값을 보여서 가장 보수적인 방법인 것으로 보인다.

APOA1, CETP와 AGT의 결과를 비교해 보면, HWE 가정 사항을 따르지 않을 때에도 스코어 방법과 H-plus 방법은 민감하지 않은 것으로 보이나 Chaplin 방법만은 영향을 많이 받는 것을 볼 수 있다. 이는 앞에서도 언급했듯이 후향적 모델의 특성때문인 것으로 보인다. 또한 로지스틱모형을 사용한 Chaplin과 H-plus방법간의 차이가 존재하는 것은 일배체를 구축하는 알고리즘의 차이 때문인 것으로 보이는데, Chaplin 방법은 한번에 일배체를 구축하는 것에 반해 H-plus는 EE 알고리즘의 Forward-block-computational 알고리즘(그림 2.1)을 이용하여 빈도가 극히 낮은 일배체는 제거하면서 순차적으로 일배체를 구축하기 때문인 것으로 보인다.

4.2.2 불완전자료에서 유전요인을 고려한 관련성 분석 결과

불완전자료에서 3개 이상의 SNP을 가진 유전자를 이용하여 고혈압과 일배체의 관련성을 분석하였고 결과는 표 4.3과 같다. 완전자료에서와 마찬가지로 유의확률에 $-log_{10}$를 취하여서 그래프로 나타내었다(그림 4.5).

표 4.3 불완전 자료에서 유전요인을 고려한 관련성 분석 결과

	SNP수	HWE pool	p-value		HWE controls	p-value	
			HTR	Score		Chaplin	H-plus
ACE	6	Yes	0.22	0.149	No	0.094	0.129
AGT	4	Yes	0.63	0.479	Yes	0.493	0.463
ALOX5AP	3	Yes	0.83	0.797	No	0.857	0.900
APOA5	5	Yes	0.03	0.071	Yes	0.537	0.220
CETP	3	Yes	0.38	0.378	No	0.384	0.629

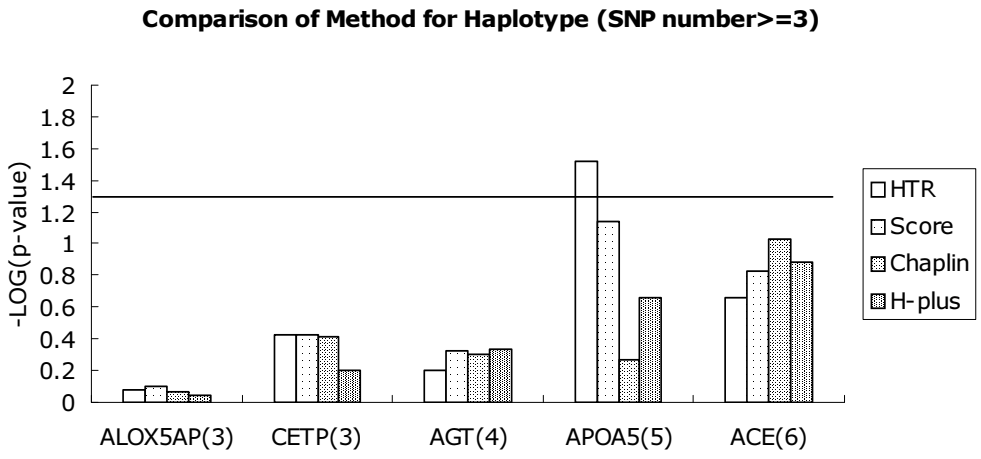

그림 4.5 불완전자료에서 3개 이상의 SNP을 이용한 관련성 분석 결과

APOA5를 봤을 때, HTR방법은 나머지 세 방법과 다르게 관련성이 있다는 결과를 보여주었다. 완전자료와 불완전자료의 비교는 다음에서 논한다.

4.2.3 완전자료와 불완전자료에서 유전요인을 고려한 관련성 분석 결과 비교

이상으로 완전자료와 불완전자료에서 유전요인만을 고려한 고혈압과 일배체의 관련성 분석을 하고 이를 비교하였다. 전체적으로 완전자료에 비해 불완전 자료는 $-log_{10}$값이 떨어지는 것으로 보였으나, APOA5에서 결측치의 수가 4임에도 불구하고 Chaplin 방법은 값의 차이가 많이 나는 것으로 나타났다. 즉, SNP수가 많은 경우 Chaplin 방법은 결측치에 민감한 것으로 보인다. HTR 방법의 경우 APOA5 유전자를 보면 완전자료에서보다 불완전자료에서 p-value가 증가하는 현상이 나타났고 이는 HTR 방법이 안정적인 방법이 아니라는 것을 알 수 있다.

각각의 유전자 별로 보면 CETP와 ACE의 경우 완전자료에 비해 불완전자료에서 $-log_{10}$값이 떨어졌는데 결측치의 비율이 CETP가 더 높은데도 불구하고 ACE에서 변동이 큰 이유는 SNP의 수가 많기 때문인 것으로 보인다. 또한 CETP만 보았을 때 SNP수도 값 저하에 영향을 미치는 것으로 보이는데 CETP는 SNP수가 3개이므로 1개의 결측치가 존재하면 2개의 SNP으로 일배체를 추정하기 때문에 차이가 심해지는 것으로 보인다.

4.3 유전요인과 환경요인을 고려한 관련성 분석 결과

4.2절은 고혈압과 일배체의 관련성을 환경요인은 고려하지 않고 분석하였다. 그러나 당뇨, 심장병, 고혈압과 같은 복합형질은 유전요인과 몸무게, 나이, 음주여부와 같은 환경요인이 유전요인과 동시에 질병과 관련성이 있다는 것이 알려져

있으며 따라서 4.3절에서는 환경적인 요인을 제어했을 때의 고혈압과 일배체의 관련성을 보고자 한다. 4.2절에서 다룬 네 가지 방법 중에서 환경요인을 다룰 수 없는 HTR 방법과 Chaplin 방법은 비교 대상에서 제외하였고, 고려해야 할 환경요인으로는 비만도, 음주 여부, 흡연 여부를 선택하였다. 비만은 BMI(Body Mass Index)값이 25이상인 사람들은 비만이 있다고 보았다. 스코어 방법과 H-plus 방법 모두 연속형의 환경요인은 다룰 수 없기 때문에 다른 환경요인들은 고려할 수 없었고 성별의 경우 APM, ALOX5AP, APOA5는 여자만 존재하였기 때문에 고려 대상에서 제외하였다.

4.3.1 완전자료에서 유전요인과 환경요인을 고려한 관련성 분석 결과

완전자료에서 환경요인을 고려한 경우와 고려하지 않은 경우를 그림 4.1과 그림 4.6을 통해 비교해 보았다. 환경요인을 고려하지 않은 경우보다 고려한 경우가 $-log_{10}$값이 크게 나왔는데, 이는 환경요인을 통제한 후 질적형질과 일배체의 관련성이 더 커졌다고 볼 수 있으며 분석하고자 하는 형질이 복합형질인 경우 유전적인 요인으로만 분석하는 것은 바람직하지 않은 것으로 보인다. 분석 결과를 보면 APOA5가 스코어 방법에 의해 유의한 검정 결과를 보였고, H-plus 방법에 의해서는 ACE가 관련성이 있는 것으로 보였다. 다른 유전자들은 고혈압과 관련성이 없음을 보여주고 있다. 전체적으로 스코어 방법과 H-plus 방법간의 결과 차이는 SNP수가 5이상일 경우를 제외하고는 없는 것으로 나타났다.

표 4.4 완전자료에서 유전요인과 환경요인을 고려한 관련성 분석 결과

	SNP수	HWE pool	p-value Score	HWE controls	p-value H-plus
ACE	6	Yes	0.074	No	0.048
AGT	4	Yes	0.374	Yes	0.375
ALOX5AP	3	Yes	0.677	No	0.839
APM	2	No	0.655	Yes	0.563
APOA1	2	No	0.534	No	0.473
APOA5	5	Yes	0.023	Yes	0.263
CETP	3	Yes	0.149	No	0.203
ESEL	2	Yes	0.413	Yes	0.587
MTP	2	Yes	0.763	Yes	0.689

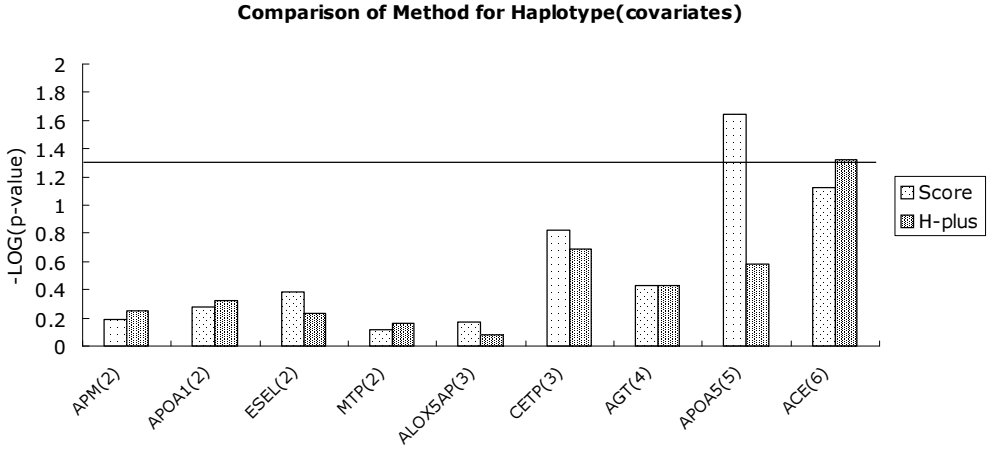

그림 4.6 완전자료에서 유전요인과 환경요인을 고려한 관련성 분석 결과

4.3.2 불완전자료에서 유전요인과 환경요인을 고려한 관련성 분석 결과

불완전자료에서 환경요인을 고려한 경우와 고려하지 않은 경우를 그림 4.4과

그림 4.7을 통해 비교해 보았다. 환경요인을 고려하지 않은 경우보다 고려한 경우가 전체적으로 $-log_{10}$값이 높게 나왔는데, 이는 완전자료와 마찬가지로 환경요인이 질병과 관련성이 있음을 보여주는 것이라고 생각되어 진다. 전체적으로 스코어 방법과 H-plus 방법간의 결과 차이는 SNP수가 5이상일 경우를 제외하고는 없는 것으로 나타났다.

표 4.5 불완전자료에서 유전요인과 환경요인을 고려한 관련성 분석 결과

	SNP수	HWE pool	p-value	HWE controls	p-value
			Score		H-plus
ACE	6	Yes	0.075	No	0.048
AGT	4	Yes	0.353	Yes	0.375
ALOX5AP	3	Yes	0.572	No	0.839
APOA5	5	Yes	0.181	Yes	0.263
CETP	3	Yes	0.302	No	0.203

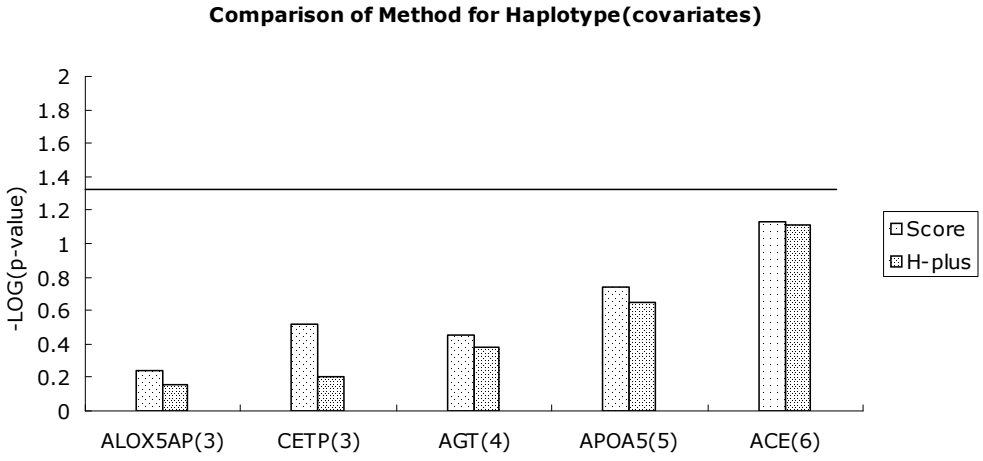

그림 4.7 불완전자료에서 유전요인과 환경요인을 고려한 관련성 분석 결과

제 5장 결론 및 토의

지금까지 선형모형에 기초한 질적형질과 일배체의 관련성을 분석하는 방법을 실재자료를 이용하여 적용, 비교하였다. 네 가지 방법은 모호한 일배체의 빈도를 추정하는 단계에서 EM 알고리즘 혹은 그의 응용방법인 ECM 알고리즘과 EE 알고리즘을 사용하였고 특정 혈질과 일배체의 관련성을 분석하는 방법에 있어서는 선형모형에 기초하였다.

본 논문에서는 고혈압과 심혈관계질환에 관련된 후보 유전자를 이용하여 실재분석을 하였으며 결과를 비교하였다. 유전요인만 고려했을 때 각 방법별로 비교를 해보면 HTR의 경우는 SNP수와 상관없이 일관되지 못한 결과를 보여주었다. HTR을 제외한 나머지 세 방법을 비교해 보면 스코어 방법이 비교적 낮은 유의확률을 보였고 Chaplin 방법은 다른 방법에 비해 높은 유의확률을 보여서 가장 보수적인 방법인 것으로 보인다. 완전자료와 불완전자료 관계없이 SNP수가 늘어날수록 방법간의 차이가 많이 나는 것으로 나타나는데 이는 SNP의 수에 따라 구성되는 일배체의 조합의 종류가 많아지기 때문인 것으로 보인다. 또한 HWE 가정 사항을 따르지 않았을 경우 방법간의 차이가 존재하였는데 이는 전향적 모델과 후향적 모델의 특성이 결과에 영향을 미치기 때문인 것으로 보인다. 환경요인을 분석에 포함한 경우와 그렇지 않은 경우를 비교한 결과 환경요인을 모형에 포함시켰을 경우 전체적으로 유의확률이 낮게 나오는 것으로 보아 환경요인이 고혈압과 관련성이 있다는 것을 알 수 있었다.

결론적으로 HTR 방법은 가장 일관성이 없었으며 스코어 방법은 가장 덜 보수적이었고 가장 보수적인 방법은 Chaplin 방법이었다. 그러나 Chaplin 방법은 SNP수가 많은 경우 결측치에 가장 민감하게 반응한다는 단점이 있다. 따라서 완전자료에서 HWE 가정 사항을 따르지 못할 경우는 Chaplin 방법보다는 스코어 방법이나 H-plus 방법을 사용하는 것이 바람직하다. 환경요인을 분석에 포함시켰을 경우 전체적으로 유의확률 값이 떨어지는 것으로 보아 일배체만 고려했을 경우 일배체와 형질과의 관련성을 더 크게 해석하는 오류를 범할 수 있음을 알 수 있었

다. 환경요인을 고려한 분석에서 사용한 스코어 방법과 H-plus방법간의 차이는 대체적으로 없는 것으로 보였다.

본 논문에서는 일배체와 환경요인의 교호작용을 보지 않았으나 추가적으로 환경요인과 일배체의 교호작용의 효과까지도 분석할 수 있는 방법에 대한 연구가 필요할 것이다. 또한 본 논문에서는 2~6개의 SNP을 이용하여 분석하였으나 실제 일배체와 형질과의 관련성을 보기 위해서는 더 많은 수의 SNP이 필요할 것으로 보인다. 따라서 SNP수의 증가로 자유도가 커지는 경우 이를 보완하는 방법에 대한 연구가 진행되어야 할 것이다.

참 고 문 헌

송 기준, 양적 형질 유전자의 연관 및 관련성에 관한 동시적 분석, 연세대학교 박사학위논문, 2003.

이 은혜, 반수체를 이용한 관련성 분석 방법 비교, 연세대학교 석사학위논문, 2004.

Boos DD. On generalized score tests. *The American Statistician*, 1992;46:327-333.

Breslow NE, Day NE. Statistical methods in cancer research. International Agency for Research on Cancer, Lyon, France, 1980.

Carroll RJ, Wang S, Wang CY. Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 1995;90:157-169.

Demster AP, Laird NM, RUbin DB. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of Statistical Society*, 1977;39:1-38.

Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotyping frequencies in a diploid population. *Molecular Biology and Evolution*, 1995;12:921-927.

Epstein MP, Satten GA. Inference on haplotype effects in case-control studies. *American juornal of human genetics*, 2003;73:1316-1329.

Kruglyak, L. and and Nickerson, D. Variation is the spice of life. *Nature*

Genetics, 2001;27:234-236.

Michalatos-Beloin S, Tishkoff S, Bentley K, Kidd K, Ruano G. Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. *Nucleic Acids Research*, 1996;24:4841-4843.

Meng X-L. Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometika*, 1993;80:267-278.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometika*, 1979;66:403-411.

Schaid DJ, Rowland CM, Tines DE, Jacobson, RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American journal of human genetics*, 2002;70:425-434.

Stephens, M., Smith, N.J. and Donnnelly Y, P. A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*, 2001;68;518-522.

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wanger MJ, Ehm MG. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human heridity*, 2002;53:79-91.

Zaho LP, Li SS, Khalid N. A method for the assessment of disease association with single-nucleotide polymorphism haplotype and environmental variables in case-control studies. *American journal of human genetics*, 2003;72:1231-1250.

ABSTRACT

A Comparison of Association Analysis using Linear Model between Qualitative trait and Haplotype

Han, Hae Ree
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

One of the recent methods for evaluating associations between genes and diseases is using SNP(single nucleotide polymorphism). Since most of diseases such as cancer, diabetes, schizophrenia, and coronary heart diseases are complex traits, methods of using single SNP have many restrictions, so the construction of haplotype which is composed of multiple SNPs within candidate genes is suggested for overcoming these restrictions. Especially, when SNPs are not independent and these are in linkage disequilibrium, haplotype-based analysis is useful in association studies.

Since phase of haplotype is not unknown, we need methods which estimate haplotype frequency to resolve this haplotype ambiguity. Nevertheless EM algorithm is most popular because of stable performance, estimation using EM algorithm may be failed when there are number of SNPs. Furthermore, ECM algorithm and EE algorithm which are applying EM algorithm are suggested.

For haplotype-based analysis we compared HTR, Score, Chaplin and H-plus

method based linear model. In addition, we considered incomplete data set which has missing data and chose models which included genetic factors and environmental factors.

In order to compare results of four methods, we used Cardiovascular genomic center data and hypertension as qualitative trait and observed some interesting results that the performance difference of four methods increased when the number of SNP was small or large. Moreover, HTR showed inconsistent results and Chaplin was considered most conservative method. When SNPs is not in HWE, Score method or H-plus method are recommended. Moreover, when we considered environmental factors, we knew environmental factors affected association of haplotype and disease and the difference of Score method and H-plus method was not showed.

Key Words : Haplotype, Association Analysis, Linear Model, HTR method, Score method, Chaplin method, H-plus method