응용통계연구 제 18권 3호, 2005년, pp. 533-541

연관분석을 위한 베이지안 모형 선택: 상호상관성 변수를 중심으로

서영주[1)]

요 약

본 저자는 앞선 연구에서 제안한 SSVS 방법을 이용하여 한 양적형질에 대한 연관분석에 있어, QTL에 가까이 있는 관련된 표지유전자들의 위치를 정하고자 한다. 본 논문에서는 QTL에 연관되어 있고 동시에 서로 연관되어 있는 몇 가지 표지유전자들을 대상으로 하는데, 이 유전자 좌위들의 i.b.d. 값들을 상호 상관이 있는 예측변수로서 고려하여, SSVS 방법으로 분석한다. 두개의 QTL에 강하게 연관되어 있는 표지유전자들 만을 동시에 고려한 분석의 결과, QTL에 가장 가까이 위치한 표지유전자가 다른 유전자들보다 더 분명하게 양적형질과의 관련성을 보여주었다. SSVS를 이용한 상호 상관이 있는 표지 유전자들의 분석의 결과는 전통적인 다중회귀분석을 이용한 결과와 거의 일치했다. 본 모의실험을 바탕으로, 복합 양적형질에 대하여 서로 연관된 다중의 표지유전자들을 동시에 연관분석을 수행하는 데에 SSVS 방법이 상당히 유용하다고 결론 내린다.

주요용어: SSVS, 연관분석, QTL.

1. 서론

가족적 형질(familial trait)에 연관된(linked) 주요유전자(major gene)를 결정하기 위하여, 형질에 대한 표지 대립유전자들(marker alleles)의 연관분석(linkage analysis)방법이 적용될 수 있다. 복합양적형질(complex quantitative trait)에 대한 연관분석은 일반적으로 형질과 표지유전자들(genetic markers) 간의 상관관계를 토대로 하여 수행된다. 널리 알려진 비모수 연관분석방법 중의 하나가 형제 쌍 연관분석(sib pair linkage analysis)이다. 이것은 형제 쌍(sib pairs)이 동일한 조상으로부터 물려받아 함께 가지고 있는 (identical by descent: 이하 i.b.d.) 대립유전자에 대한 형질의 특성을 측정하는 방법이다. 형제 쌍 연관분석 기법으로서 일반적으로 널리 알려진 것은 Haseman-Elston의 방법(1972)으로서, 한 표지유전자 좌위(locus)에서의 형제 쌍의 i.b.d. 대립유전자 수와 양적형질을 결정하는 가정된 양적형질 유전자좌위(quantitative trait locus: QTL)에서의 i.b.d. 수 사이의 재조합률(recombination fraction)과 관련한 결합분포를 이용하여, 주어진 표지유전자에서 형제 쌍의 i.b.d. 대립유전자 수를 형제 쌍 형질의 차이의 제곱으로 회귀분석하는 방법이다. 이후, Elston et al.(2000)은 종속변수에 형제 쌍의 공분산을 고려하여, 주어진 유전자 좌위에서 형제 쌍의 i.b.d. 대립유전자 수를 형질의 평균 수정한 직적(mean-corrected cross-product)으로 회귀분석하여 연관

1) (140-742) 서울시 용산구 청파동 2가 효창원길 52, 숙명여자대학교 자연과학연구소, 학술연구교수
E-mail: ysprite@hotmail.com

분석의 검정력을 높이는 개선된 Haseman-Elston의 방법을 제안하였다. 주어진 유전자 좌위에서 둘 이상의 유전자들과 관련이 되는 복합형질의 연관분석을 위해서는, 이 방법을 다중의 유전자 좌위에서의 연관(linkage)이 동시에 고려될 수 있도록 다중회귀분석 방법으로 확장시켜 적용할 수 있다.

하지만, 전통적인 다중회귀분석 방법들의 이용 시, 많은 표지유전자 집합의 분석에서 후보 표지유전자들의 모든 가능한 조합들 중 최적의 부분집합을 찾는 데에는 계산적인 부담이 있을 수 있다. 이러한 이유로 인해, 본 연구에서는 George와 McCulloch(1993)의 베이지안 회귀모형에 근거한 확률론적 탐색에 의한 변수선택(stochastic search variable selction: 이하 SSVS) 방법을 제안한다. 이 방법은 최고의 사후확률값(posterior probability)을 가진 예측변수들의 부분집합을 최적의 부분집합으로 간주하려는 것이다. 본 저자의 앞선 연구(Suh et al., 2001, 2003)에서는 SSVS에 대한 George와 McCulloch(1993)의 기법을 이용하여 연관된 표지유전자들의 최적의 부분집합을 선택하기 위한 방법을 제안하였다. 하지만 2001년의 연구에서는 한 염색체 당 하나의 표지유전자 만을 고려하였다. 따라서 고려된 표지유전자 좌위에서의 형제 쌍의 i.b.d. 값들은 서로 독립적이었다. 이 방법은 12번째 Genetic Analysis Workshop[2](GAW12: Wijsman et al., 2001)을 위한 모의 실험자료에서 Q1과 Q2의 두 양적 형질들에 대한 주요 유전자들을 포함한 염색체들의 위치를 정하는 데에 적용되었다.

본 연구에서는 QTL에 가장 가까이 있는 표지유전자들의 위치를 정하는 데에 초점을 두고자 한다. 주요 유전자 1(MG1)과 주요 유전자 2(MG2)에 연관되어 있고, 동시에 서로 연관되어 있는 표지유전자들을 대상으로 한다. MG1과 MG2는 둘 다 QTL이므로, 이제부터는 MG1과 MG2 대신, 각각 QTL1 과 QTL2로 명명한다. 연관분석을 위해서는 개선된 Haseman-Elston 방법(Elston et al., 2000)을 적용하고자 한다. 이 방법은 앞서 설명한 바와 같이 주어진 표지 유전자에서 형제 쌍의 i.b.d. 대립유전자 수를 한 형질의 평균 수정한 직적(이하 $C(Q)$)으로 회귀분석하려는 것이다. 이 방법은 서로 상관이 있는 몇 가지 표지유전자들을 대상으로 할 때, 형제 쌍의 i.b.d. 대립유전자 수에 관한 형질의 평균 수정한 직적의 다중회귀분석으로도 잘 확장될 수 있다. 본 논문의 목적은 몇 가지 상호 상관이 있는 예측변수들이 있는 경우, George와 McCulloch(1993)의 SSVS 기법을 Elston et al.(2000)의 연관분석에 적용하여 잘 수행되는지를 평가하려는 것이다. 이 방법은 연관분석에서 QTL에 가장 가까이 있는 표지유전자들의 위치를 정하고자 하는 상황에 적용될 것이다. 또한 이 방법은 알고 있는 한 후보유전자 좌위에서, 이 부위의 단일염기다형성(Single Neucleotide Polymorphism: SNP) 분석의 결과들을 이용하는 상황에도 적용될 수 있을 것이다.

2. 자료 및 분석방법

본 연구에서는 GAW12에서 주어진 모의자료(problem 2)를 이용하여 분석하였다. 이는 50개의 같은 분포의 반복된 자료로서 각각 1,497명의 개인들과 23개의 확장된 가계(extended pedigree)들로 구성되어 있다. 각 개인의 22개의 각 상염색체 상에는 58에서 220개(총 2,855개)의 유전자들이 위치하고 있다. 7개의 주요 유전자 요인(genetic factor)와 성, 연령과 환경적

2) GAW는 통계유전학적 방법의 연구 및 비교 평가를 목적으로 하는 공동연구회로서 매 짝수년도에 개최된다. GAW12는 2000년 10월 23-26일 미국 Texas주 San Antonio에서 있었다.

요소 같은 비 유전적 요인(non-genetic factor)에 각기 다른 형태로 영향을 받는 5개의 양적 형질들은 특정 병의 발병에 영향을 준다. 본 연구에서는 Q1로 명명한 한 양적형질에 대하여 주요 유전자들에 가장 가까이 있는 표지유전자들의 위치를 정하기 위하여 연관분석을 수행하였다. 이 형질은 19번째 염색체 상의 QTL1(MG1)과 두 번째 염색체 상의 QTL2(MG2)의 두 개의 주요 유전자들과 성과 연령을 포함한 비 유전적 요인들에 의해 결정된다. 이 비 유전적 요인의 공변량들로 보정 하에, 유전적 분산성분모형(genetic variance component model)에서는 QTL1이 Q1에서의 분산의 24%를, QTL2가 21%를 차지한다. 전체 50개의 자료에는 표본 당 771개의 형제 쌍이 포함되어 있는데, 이 표본들을 각각 따로 분석하기에는 검정력이 매우 낮았다. 결국 연관분석에서의 타당한 검정력을 얻기 위해, Suh et al.(2001, 2003)에서 수행한 바와 같이, 50개의 모의자료를 $5k-4(k=1,2,\cdots,10)$부터 $5k$까지 5개씩 결합하여 각 표본 당 $n=3,855$개의 형제 쌍을 포함한 10개의 대 표본들로 구성하였다.

Suh et al.(2001, 2003)에서는 수정된 Haseman-Elston 통계량(Elston et al., 2000)을 이용하여 분석하였다. 종속변수는 가계 내에서 j번째 형제 쌍에 대한 Q1의 평균 보정한 직적으로 다음과 같이 정의된다.

$$C(Q1_j) = (Q1_{j1} - m)(Q1_{j2} - m)$$

여기서 $Q1_{j1}$와 $Q1_{j2}$는 j번째 형제 쌍에서 첫 번째와 두 번째 형제 각각에 대한 형질의 관찰된 형질 값들이고, m은 모든 i와 j에 대한 $Q1_{ij}$의 평균이다. Q1을 연령과 성으로 보정한 값을 $Q1^*$로 명명하고, 이에 대한 연관분석을 하였다. $Q1^*$에 대한 형질 직적인 $C(Q1^*)$가 모형에서 고려되었다.

$n=3,855$개의 형제 쌍을 각각 포함한 10개의 표본들에 대하여, QTL1에 가까운 6개의 표지유전자들과 멀리 떨어진 6개의 표지유전자들과, QTL2에 가까운 6개의 표지유전자들과 멀리 떨어진 6개의 표지유전자들을 각각 가능한 인자들로 고려하여 분석을 수행하였다. 관련한 예측변수들로는 (1) QTL1에 가장 가까이 위치한 표지유전자(I_{11}), QTL1에 가까운 다른 6개의 표지유전자들(I_{1m}, $m=2,\cdots,7$)과, QTL1과 같은 염색체 상에서 멀리 위치한 6개의 표지유전자들에서 형제 쌍의 i.b.d. 대립유전자 수의 표준화된 값들(I_{1m}, $m=8,9,\cdots,13$), 그리고 (2) QTL2에 가장 가까이 위치한 표지유전자(I_{21}), QTL2에 가까운 다른 6개의 표지유전자들(I_{2m}, $m=2,\cdots,7$)과, QTL2와 같은 염색체 상에서 멀리 위치한 6개의 표지유전자들에서 형제 쌍의 i.b.d. 대립유전자 수의 표준화된 값들(I_{2m}, $m=8,9,\cdots,13$)이 고려되었다. SSVS 기법의 결과들은 빈도론자 단순회귀분석 및 다중회귀분석의 결과들과 비교되었다. $Q1^*$에 대한 단순회귀분석에서는 각 표본 당 26번($n=26$)의, 즉 한 표지유전자 당 한 번의 회귀분석이 수행되었다. 다중회귀분석은 각 표본마다 QTL1과 QTL2 각각에 대해서 독립적으로 두 번 수행하였다(매 분석 당 13개의 표지유전자가 고려됨). 단순회귀분석의 분석식은 $C(Q1^*)=\alpha+\beta_{im}I_{im}+\epsilon$이며, I_{im}은 QTLi($i=1,2$)에 대한 유전자 좌위를 가지고 있는 염색체 상의 m($m=1,2,\cdots,13$)번째 표지유전자 단변량에 대한 형제 쌍 i.b.d. 대립유전자 수의 표준화된 값이다. 또한, 다중회귀분석에서는 먼저 QTLi($i=1,2$)에 대하여 $C(Q1^*)=\alpha+\sum_{m=1}^{13}\beta_{im}I_{im}+\epsilon$을 고려하였다. 다음으로, QTL1과 QTL2에 관련한 표지유전자들을 동시에 고려한 모형을 분석하였다. 이 분석을 위하여, 26개의 후보 예측유전자들 $I_{im}(i=1,2;\ m=1,2,\cdots,13)$을 회귀모형에서 다음과 같이 고려하

였다: $C(Q1^*) = \alpha + \sum_{m=1}^{13} \beta_{1m} I_{1m} + \sum_{m=1}^{13} \beta_{2m} I_{2m} + \epsilon$. 단순회귀분석에서는 고려한 두 개의 염색체들 각각에 대해 최대 T값(maximum T value)을 가진 표지유전자를 선택하였다. 다중회귀분석에서는 $T \geq 2.14$(즉, LOD score ≥ 1.0)를 기각역으로 하였다. 모든 변수들을 포함하여 한번에 입력 후 후진제거법 및 단계적 전진선택법으로 다중회귀분석을 수행하였다.

George와 McCulloch(1993)가 제안한 SSVS 기법은 Suh et al.(2001, 2003)에서와 같이 수정된 Haseman-Elston 방법(2000)을 이용한 연관분석에 적용되었다. 확장된 방법은 모든 p개 후보 표지유전자들을 포함한 다중회귀모형과 관련한다. 본 연구 분석에서는 $Q1^*$에 대하여 QTL1과 QTL2 각각에 대한 모형에서 $p = 13$개의 예측변수들을 고려하였다. 비슷하게, QTL1과 QTL2에 관련한 표지유전자들을 동시에 분석하는 모형에서는 $p = 26$개의 예측변수들을 고려하였다. β_i의 사전확률분포는 $\beta_i|\gamma_i \sim (1-\gamma_i)N(0,\tau^2) + \gamma_i N(0, c^2\tau^2)$의 정규혼합분포이며, 여기서 $\gamma_i \sim Bernoulli(1/2)$, $i = 1, 2, \cdots, p$이다. σ^2의 사전확률분포는 $\sigma^2 \sim IG(0.001, 0.001)$인데, IG는 역감마분포를 의미한다. 본 연구의 사전분포에서 τ와 c에 대하여 $(\sigma_{\beta_i}/\tau, c) = (10, 100)$의 반자동적 값을 고려하였는데, 여기서 σ_{β_i}는 단변량 모형에서 β_i의 표준오차이다. 이 방법은 τ와 c의 값과 관련된 $\gamma_i (i = 1, 2, \cdots, p)$의 사후분포에 초점을 둔다는 점에서 다른 베이지안 방법과 구분된다. 깁스표본추출을 이용한 SSVS 분석이 수행되어, p개의 예측변수 I_i에 대한 $\underline{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$의 사후분포가 생성된다. 여기서의 주 관심사는 최고의 사후확률을 갖는 $\underline{\gamma}$의 값, 즉 $\underline{\gamma}^+ = (\gamma_1^+, \gamma_2^+, \ldots, \gamma_p^+)$와, $\gamma_i^+ = 1$인 표본들의 수를 구하는 것이다. $C(Q1^*)$를 예측하는 최적의 모형은 $\gamma_i^+ = 1$에 대응하는 I_i들로 구성된다.

본 연구에서는 회귀모형을 분석하기 위해 SAS를 이용하였다. 형제 쌍의 연관분석을 위해서는 C 프로그램이 실행되었다. SSVS를 수행하는 데에는 WinBUGS가 이용되었는데, 실행시 각 표본 마다 1,000번의 반복 소거(burn-in) 후 5,000번의 반복 수행을 하였다.

3. 분석결과

먼저, 두 표지유전자들 간의 물리적인 근접이 각 유전자의 형제 쌍 i.b.d. 대립유전자 수의 공선형성에 어느 정도 영향을 미치는지를 알아보기 위해, 표지유전자 변수들 사이의 상관관계를 분석하였다. QTL1과 QTL2 각각에 대하여, 이 주요 유전자들에 강하게 연관되어 있는 7개의 표지유전자들과 QTL로부터 멀리 떨어져 있는 6개의 표지유전자들의 형제 쌍 i.b.d. 대립유전자 수들 간의 상관계수($r_j, j = 1, \cdots, 78$)를 표지유전자들의 각 쌍들 사이의 유전자지도 거리(genetic map distance; $m_j, j = 1, \cdots, 78$)의 함수로서 관계를 살펴보았다. i.b.d. 수에 대한 추정된 상관계수 값(r_j)들과 유전자지도 거리 값(m_j)들 간에 유의한($P < 0.01$) 음의 상관관계를 얻었는데, QTL1과 QTL2에 대한 추정 상관계수는 $n = 78$개의 표지유전자 쌍에 대하여 각각 $r = -0.76$과 $r = -0.99$였다. 이 결과로서 같은 염색체 상에서 서로 밀접하게 위치할수록 표지유전자들의 형제 쌍 i.b.d. 대립유전자 수들 간에 강한 상관관계가 있음을 보여주었다. 반면에, 멀리 떨어져 위치한(30cM 이상, 즉 재조합률 $\theta > 0.3$) 표지유전자들의 형제 쌍 i.b.d. 대립유전자 수들 간에는 상관관계가 거의 없는 것

으로 나타났다($|r_j| < 0.1$).

예비단계로서, 50개의 표본을 결합한 전체 "인구"($n = 38,550$의 형제 쌍을 포함)를 이용하여 QTL1과 QTL2 각각에 연관되어 있는 표지유전자들에 대해서 $C(Q1^*)$에 관한 단계적 회귀분석을 수행하였다. QTL1의 경우, $C(Q1^*)$의 분석에 있어 I_{13}(D19G030), I_{14}(D19G031)와 I_{15}(D19G033)가 $T \geq 2.14$(즉, LOD score ≥ 1.0) 수준에서 유의한 예측변수들이었는데, 이 때 $R^2 = 0.0051$이었다. 하지만, QTL1에 가장 가까이 위치한 표지유전자 변수인 I_{11}(D19G032)이 회귀모형에서 유의하지 않았던(단순회귀분석에서는 $P < 0.0001$ 이지만, 다중회귀분석을 이용해서는 $P = 0.16$) 것은 흥미로웠다. QTL2의 경우, QTL2에 가장 가까이 위치한 표지유전자의 i.b.d. 값인 I_{21}(D02G170)과 I_{24}(D02G169)가 $T \geq 2.14$의 임계값에서 유의한 예측변수들로 선택되었다. 다음으로, QTL1과 QTL2 각각에 가까이 위치한 표지유전자(QTL1과 QTL2에 가장 가까이 위치한 표지유전자를 포함하여) 변수들만을 한번에 고려하였다. 즉, QTL1에 가까이 있는 6개의 표지유전자들에서의 i.b.d. 값들과 I_{11} 및 QTL2에 가까이 있는 6개의 표지유전자들에서의 i.b.d. 값들과 I_{21}을 후보 예측변수들로서 고려하였다. $n = 38,550$의 형제 쌍으로 이루어진 결합한 표본에 대하여, $T \geq 2.14$의 임계값을 이용하여 분석한 결과, QTL1에 관련한 I_{13}(D19G030), I_{14}(D19G031)와 I_{15}(D19G033) 및 QTL2에 관련한 I_{21}(D02G170)과 I_{24}(D02G169)의 5개의 예측변수들이 유의하게 선택되었는데, 이는 앞선 결과와도 같은 것이었다.

본 연구에서는 첫 번째로, 같은 염색체 상에 있는 몇 가지 표지유전자들을 한 형질에 대한 주요 유전자로서 관찰하고자 하였다. 이 표지유전자 변수들은 상관관계가 있는데, 특히 염색체 상에서 서로 가까이 위치하고 있는 표지유전자 변수들은 더욱 관련이 있다. 19번째 염색체상의 표지유전자들의 경우(QTL1), SSVS와 다중회귀분석(후진제거법을 이용)의 결과, 단순회귀분석에서 최대 T값을 가져 선택된 변수(이 경우 표지유전자 좌위)를 거의 포함하였다. QTL1로부터 멀리 떨어진(41cM 이상) 6개의 표지유전자들 중 하나가 SSVS 또는 후진제거법을 이용한 다중회귀분석를 통하여 10개의 대 표본들 가운데 하나의 표본에서 탐지되었다(제 1종 오류 = 1/10). QTL1에 대한 SSVS의 결과는 표본의 70%에서 단순회귀분석의 결과와 같게 얻어졌다. 두 번째 염색체상의 표지유전자들에 대해서는(QTL2), QTL2로부터 멀리 떨어진(157cM 이상) 6개의 표지유전자들 모두가 SSVS 또는 후진제거법을 이용한 다중회귀분석의 결과, 유의하지 않게 나타났다(제 1종 오류 = 0/10). QTL2에 대한 SSVS의 결과는 다중 및 단순회귀분석의 결과와 100% 같았다.

두 번째로, QTL1과 QTL2에 관련한 표지유전자들을 동시에 고려하였다. 즉, QTL1과 QTL2 각각에 가장 가까이 위치한 표지유전자, QTL1과 QTL2 각각에 강하게 연관되어 있는 다른 6개의 표지유전자들과, QTL1과 QTL2 각각과 같은 염색체 상에서 멀리 떨어져 있는 6개의 표지유전자들을 동시에 분석하였다. 본 연구에서는 표 3.1에서와 같이, (1) SSVS와 (2) 단계적 전진선택 및 후진제거법을 이용한 다중회귀분석의 결과를 비교하였다. 여기에서는 10개의 대 표본 각각에서 복수의 표지유전자들이 QTL에 연관된 표지유전자들로서 탐지되었을 때, 탐지된 표지유전자들 가운데에서 QTL1과 QTL2 각각에 가장 가까이 위치한 표지유전자 하나씩 만을 선택하였다. 이는 QTL로부터 최단 몇 cM 내에서 QT를 설명할 수 있는 표지유전자가 탐지되는지에 주 관심을 두었기 때문이다. 이 결과는 QTL2에 가

표 3.1: 탐지된 표지유전자 좌위로부터 QTL까지의 최단거리에 대한 비율 비교

			거리 내에서 표지유전자를 탐지하는 표본들의 비율	
			SSVS[a]	다중회귀[b]
QTL1로부터의 거리[c]	2.13	[D19G032][d]	1/10[e]	1/10
	2.47	[D19G031]	4/10	4/10
	4.48	[D19G033]	4/10	4/10
	5.73	[D19G030]	0/10	0/10
	6.30	[D19G035]	0/10	0/10
	6.38	[D19G036]	1/10	1/10
	7.33	[D19G028]	0/10	0/10
	39.59	[D19G003]	0/10	0/10
	40.68	[D19G002]	0/10	0/10
	41.40	[D19G001]	0/10	0/10
	68.72	[D19G102]	0/10	0/10
	68.99	[D19G103]	0/10	0/10
	70.14	[D19G104]	0/10	0/10
QTL2로부터의 거리	0.74	[D02G170]	4/10	5/10
	1.41	[D02G169]	2/10	1/10
	1.62	[D02G168]	1/10	1/10
	1.91	[D02G172]	1/10	2/10
	2.39	[D02G167]	0/10	0/10
	2.66	[D02G173]	1/10	1/10
	4.01	[D02G174]	1/10	0/10
	157.77	[D02G012]	0/10	0/10
	159.78	[D02G011]	0/10	0/10
	159.79	[D02G010]	0/10	0/10
	160.78	[D02G009]	0/10	0/10
	161.49	[D02G006]	0/10	0/10
	163.08	[D02G004]	0/10	0/10

[a] 주어진 거리 이내에 있는 표지유전자가 SSVS에 의해 유의하게 선택된 부분집합에 포함된 표본의 비율.

[b] 단계적 전진선택 및 후진제거법을 이용한 다중회귀분석: 주어진 거리 이내에 있는 표지유전자에 대하여 $T \geq 2.14$, 즉 LOD score ≥ 1.0으로 유의하게 관측된 표본의 비율.

[c] QTL1(위치: 42.1cM) 또는 QTL2(위치: 173.2cM)로부터의 유전자지도 거리(cM).

[d] 표지유전자 수(marker number).

[e] 각 표본 당 3,855개의 형제 쌍을 포함한 10개의 대표본.

장 가까이 위치한 표지유전자(0.74cM 이내에 있는 D02G170)가 다른 표지유전자들보다 더 강하게 연관되어 있음을 지지하는 분명한 증거를 보여준다(표 3.1 참조). 이 경우, 4cM 이내에 있는 연관된 표지유전자들이 10개의 대 표본들에서 선택되었다. QTL1에 대해서는, 5cM 이내에 위치한 연관된 표지유전자들 모두가 SSVS 또는 단계적 전진선택법을 이용한 다중회귀분석을 통하여 탐지되었다. 하지만, QTL1에 가장 가까이 위치한 연관된 표지유전자(QTL1의 2.13cM 이내에 있는 D19G032)는 표본들 중 1/10에서만 탐지되었다. SSVS 기법으로써 얻어진 결과는 후진제거법(혹은 단계적 전진법)을 이용한 다중회귀분석을 통해 얻어진 결과와 거의 일치하였다.

4. 토의 및 결론

베이지안 최적 부분집합 회귀에 근거한 SSVS 기법은 모든 후보 모형들 가운데에서 최고의 사후확률에 근거한 최적의 부분집합을 선택하려는 것이다. 이 방법은 모든 가능한 모형들을 고려하는 "빈도론자" 최적의 모형선택방법과 비교하여 계산적 부담을 줄인다. 하나의 염색체 당 하나의 표지유전자 만을 고려하여, 관련한 i.b.d. 값들이 서로 독립인 경우, Q1 유전자 좌위를 포함하는 염색체를 확인하는 검정력은 Suh et al.(2001, 2003)에서 보고한 바 있다. 본 논문에서는 QTL에 연관될 뿐만 아니라 서로 연관된 몇 가지 표지유전자들을 고려하였다. George와 McCulloch(1993)의 SSVS 기법과 전통적인 회귀분석 방법을 이용하여 몇 가지의 공선형성이 있는 예측변수들을 연구하였다. 비록 QTL에 가장 밀접하게 위치해 있는 표지유전자를 탐색하기 위한 연관분석의 검정력이 일반적으로 높지는 않았지만, 형질 표지유전자에 연관된 표지유전자들을 확인하기 위하여 SSVS 기법을 이용한 결과는 Elston et al.(2000)의 방법을 응용한 다중회귀분석의 결과와 거의 같았다.

이상과 같이, 다른 유전자들과 상호 관련이 있는 몇 가지 표지유전자들을 다룰 수 있도록 하는 SSVS의 베이지안 특성에 주안점을 두었다. Suh et al.(2003)에서 언급했던 것처럼, 더욱 많은 비용의 분석을 고찰하기 이전에 공선형성이 있는 표지유전자들을 가진 최적 조건에서 SSVS가 어떻게 잘 수행되는지를 먼저 확인하는 것은 의미가 있다고 본다. SSVS 기법은 본 논문에서 수행한 바와 같이 몇 개의 인자들 보다는 더 많은 수의 인자들을 이용할 때 더 좋은 결과를 얻을 것이다. 이에 더하여, 이 방법은 적은 수의 관측 대상에 비해 많은 수의 변수들이 있을 때에도 유효하다. SSVS 방법은 Oh et al.(2003)에서 수행된 바와 같이, 많은 수의 표지유전자들을 고려하는 보다 현실적인 상황에서 효과적일 것이다. 앞으로의 연구에서는 Chipman(1996)이 제안하였고 Oh et al.(2003)에서 보여 주었듯이, SSVS에서 계층적 모형 사전분포를 고려할 것이다. SSVS 기법에서 사전분포에 상관구조가 포함될 때 더 좋은 결과가 얻어질 것이다. 덧붙여, 많은 수의 유전자 변수들을 다루고자 할 때, 본 SSVS에서 사용된 깁스표본기법 대신에 메트로폴리스-헤스팅스(Metropolis-Hastings) 알고리즘을 적용할 수 있다. Yoon et al.(2004)은 메트로폴리스-헤스팅스 알고리즘을 이용하여 사전분포 값에 약간의 변화를 준, 좀더 간단하고 효과적인 베이지안 변수선택방법을 제안하였다.

본 모의실험 연구를 바탕으로, SSVS는 복합양적형질에 대하여 다중의 연관된 표지유전

자들을 동시에 고려하는 연관분석을 수행하는 데에 상당히 유용하다고 결론 내린다. QTL이 분산의 더 큰 비율을 설명하는 연관분석에서, SSVS 기법은 정확한 염색체의 위치를 확인하고 QTL에 가까이 위치한 후보유전자들을 위치시키는 데에 더욱더 유용할 것으로 기대한다. 즉, 대립유전자 i.b.d.의 수들과 형질 공유의 정도 사이에 더 강한 관련이 있을 때, 이 방법은 더 잘 수행될 것이다. 또한 사전분포에 표지유전자 상관구조가 포함되는 경우 SSVS를 통하여 더 나은 결과를 얻으리라 생각된다.

참고문헌

Chipman, H. (1996). Bayesian variable selection with related predictors, *Canadian Journal of Statistics*, **24**, 17-36.

Elston, R. C., Buxbaum, S., Jacobs, K. B. and Olson, J. M. (2000). Haseman and Elston revisited, *Genetic Epidemiology*, **19**, 1-17.

Forrest, W. F. and Feingold, E. (2000). Composite statistics for QTL mapping with moderately discordant sibling pairs, *American Journal of Human Genetics*, **66**, 1642-1660.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association*, **88**, 881-889.

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.

Oh, C., Ye, K. Q., He, Q. and Mendell, N. R. (2003). Locating disease genes using Bayesian variable selection with the Haseman-Elston method, *BMC Genetics*, **4**(Suppl), S69.

Shete, S., Jacobs, K. B. and Elston, R. C. (2003). Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences, *Human Heredity*, **55**, 79-85.

Suh, Y. J., Finch, S. J. and Mendell, N. R. (2001). Application of a Bayesian method for optimal subset regression to linkage analysis of Q1 and Q2, *Genetic Epidemiology*, **21**(Suppl 1), S706-S711.

Suh, Y. J., Ye, K. Q. and Mendell, N. R. (2003). A method for evaluating the results of Bayesian model selection: application to linkage analyses of attributes determined by two or more genes, *Human Heredity*, **55**, 147-152.

Wang, D., Lin, S., Cheng, R., Gao, X. and Wright, F. A. (2001). Transformation of sib-pair values for the Haseman-Elston method, *American Journal of Human Genetics*, **68**, 1238-1249.

Wijsman, E. M., Almasy, L., Amos, C. I., Borecki, I., Falk, C. T., King, T. M., Martinez, M. M., Meyers, D., Neuman, R., Olson, J. M., Rich, S., Spence, M. A., Thomas, D. C., Vieland, V. J., Witte, J. S. and MacCluer, J. W. (2001). Analysis of complex genetic traits: applications to asthma and simulated data, *Genetic Epidemiology*, **21**(Suppl 1), S1-S853.

Xu, X., Weiss, S., Xu, X. and Wei, L. J. (2000). A unified Haseman-Elston method for testing linkage with quantitative traits, *American Journal of Human Genetics*, **67**, 1025-1028.

Yoon, S., Suh, Y. J., Mendell, N. R., Ye, K. Q. (2005). A Bayesian approach for applying Haseman-Elston methods, *BMC Genetics*, (in press).

[2005년 2월 접수, 2005년 5월 채택]

Bayesian Model Selection for Linkage Analyses: Considering Collinear Predictors

Young Ju Suh [1)]

ABSTRACT

We identify the correct chromosome and locate the corresponding markers close to the QTL in the linkage analysis of a quantitative trait by using the SSVS method. We consider several markers linked to the QTL, as well as to each other and thus the i.b.d. values at these loci generate collinear predictors to be evaluated when using the SSVS approach. The results on considering only closely linked markers to two QTL simultaneously showed clear evidence in favor of the closest marker to the QTL considered over other markers. The results of the analysis of collinear markers with SSVS showed high concordance to those obtained using traditional multiple regression. We conclude based on this simulation study that the SSVS is quite useful to identify linkage with multiple linked markers simultaneously for a complex quantitative trait.

Keywords: QTL; Linkage analysis; SSVS.

1) Research Professor, The institute of Natural Sciences, Sookmyung Women's University, Hyochangwon-gil 52, Chungpa-dong 2 ga, Yongsan-gu, Seoul, 140-742, Korea
E-mail: ysprite@hotmail.com