응용통계연구 제 16권 2호, 2003년, pp. 195–202

EM 알고리듬을 이용한 단일염기변이(SNP; Single Nucleotide Polymorphism)군의 일배체형(Haplotype) 비율 추정 *

김선우[1] 김종원[2] 이경아[3]

요 약

복합성유전질환 연구에 있어서 단일염기변이를 이용한 일배체형 분석은 개별적인 단일염기변이 분석에 비하여 비용 및 효율 면에서 훨씬 유용하며, 생물학적으로도 기능적 중요성을 갖는 것으로 평가되고 있다. 그러나 일반적인 유전형분석방법을 이용한 단일염기변이군 자료는 이배체형(diploid)으로서 위상(phase)을 확인할 수 없으므로 일배체형 비율을 예측하기 어렵다. 본 연구에서는 고형종양 환자군과 정상군의 단일염기변이군 이배체형 자료가 주어졌을 때 단일염기변이군 일배체형 비율의 우도함수에 EM 알고리듬을 적용하여 각 일배체형의 비율을 추정하였다. 이로부터 단일염기변이간의 연관불균형(linkage disequilibrium)을 분석하여 고형종양과 연관 가능성이 있는 단일염기변이를 살펴보았다.

주요용어: 단일염기변이, 이배체형, 일배체형, EM 알고리듬, 연관불균형.

1. 서 론

인종간 또는 개개인이 특정 질병에 대한 소인에 차이가 있고, 약물에 대한 반응성과 효과에 차이를 보이는 이유는 게놈상에 나타나는 변이로서 설명되고 있다. 이 중 가장 흔히 나타나는 변이인 단일염기다형성은 인간게놈의 100-300 염기마다 한 개 정도의 빈도로 매우 빈번하게 관찰된다. 이러한 단일염기변이의 연구는 특정 질병에 대한 민감성 진단 및 맞춤식 치료제 개발 등이 가능하므로 세계 각국은 많은 연구비를 투입하여 단일염기변이 데이터베이스를 구축하고 있다. 그러나 단일염기변이를 이용한 질병 연관성 연구는 고비용, 저효율 등의 문제점으로 제한적으로 시행되고 있다. 이러한 한계점을 극복하기 위해 단일염기변이군의 일배체형에 대한 분석이 시도되고 있다. 단일염기변이군의 gametic phase인 일배체형은 함께 유전되는 경향이 있어 유전적 변이 정보를 종합적으로 밝혀내기가 용이

* 본 연구는 과학기술부 국가지정연구실 M10203000038-02J000002110 사업의 지원에 의하여 이루어진 것임.

1) (135-230) 서울시 강남구 일원동 50, 삼성생명과학연구소, 통계지원팀
E-mail : kyuwon@samsung.co.kr
2) (135-230) 서울시 강남구 일원동 50, 성균관대학교 의과대학, 삼성의료원, 진단검사의학과
E-mail : jwonk@smc.samsung.co.kr
3) (136-705) 서울시 성북구 안암동 5가 126-1, 고려대학교 의과대학, 고려대학교 안암병원, 진단검사의학과
E-mail : kal1119@hanmail.net

하다. 즉 일배체형 분석은 집단간 유전적 유사성, 연관불균형 정도 및 단일염기변이들의 상호 기능적 관련성에 근거한 질병연관성 분석 등에 활용되고 있다. 그러나 대부분의 유전형 분석 방법은 염색체 위상을 구분할 수 없으므로 여러 개의 단일염기변이 부위로부터 얻은 유전형 분석 결과가 이형접합자(heterozygote)인 경우 직접적으로 일배체형을 알 수가 없다. 이배체형 자료로부터 일배체형을 분석할 수 있는 방법으로는 long-range PCR(Newton et al., 1989; Wu et al., 1989; Ruano et al., 1990)이나 염색체 분리와 같은 직접적 실험적 기술의 적용 및 가계조사 방법을 들 수 있다(Perlin et al., 1994). 그러나 이러한 방법은 분석을 위하여 많은 노동력 및 고비용이 요구되며, 검체수집의 어려움이 동반되므로 대량의 자료를 분석하는데 어려움이 있다. 단일염기변이 자료에 대한 통계적 방법의 적용은 대립유전자(allele) 빈도나 일배체형 비율의 추정을 가능케 한다. 그러나 대립유전자의 수를 세어 각 유전자좌의 대립유전자 비율을 추정하는 것과 같은 해석적인 최대우도추정법(Gart and Nam, 1984)은 단일염기변이군의 일배체형 비율의 추정에는 사용이 불가능하다. 그 이유는 단일염기변이의 이형접합성(heterozygosity) 때문에 이배체형의 종류보다 추정하려는 일배체형의 종류가 더 많기 때문이다. 이 경우 수치적 방법으로 우도함수를 최대화하는 일배체형의 비율을 추정할 수 있을 것이다. 그러나 단일염기변이의 수가 조금만 늘어나도 가능한 일배체형의 수가 기하급수적으로 증가하기 때문에 일반적인 수치적 방법은 사용이 어렵게 된다. 따라서 불완전 자료(incomplete data)에서 효과적으로 모수를 추정할 수 있는 EM 알고리듬(Dempster et al., 1977)을 적용함으로써 일배체형 비율을 추정할 수 있으며(Long et al., 1995; Excoffier and Slatkin, 1995), 추정된 일배체형 비율은 단일염기변이간 연관불균형이 존재하는지에 대한 가설에 대해 우도비 통계량을 통한 검정을 가능케 할 것이다. 본 연구에서는 고형종양 환자군과 정상군에서 특정 단일염기변이군에 대한 일배체형 분석을 통하여 유전적 변이를 알아보고, 고형종양과의 연관 가능성을 분석하였다. 관측된 이배체형 자료가 주어졌을 때 일배체형 비율에 대한 우도함수에 EM 알고리듬을 적용하여 고형종양 환자군과 정상군에서 일배체형 비율들을 추정하였으며, 이 자료를 활용하여 단일염기변이간 연관불균형 분석을 수행함으로써 고형종양과 연관 가능성이 있는 단일염기변이들을 알아보았다.

2. 일배체형 비율의 우도함수 및 EM 알고리듬 적용

일배체형을 알지 못하는 여러 개의 단일염기변이들이 있고, 각 단일염기변이는 이배체형으로 측정된 경우, 이배체형 자료가 주어졌을 때 일배체형 비율에 대한 우도함수 및 EM 알고리듬 적용 절차는 다음과 같다(Long et al., 1995; Excoffier and Slatkin, 1995). 관측된 이배체형을 구성하는 일배체형의 조합의 수는 이형성인 단일염기변이의 수로 결정된다. j번째 이배체형을 구성하는 일배체형의 조합이 c_j개 있고 이형성인 단일염기변이가 s_j개 있을 때 $(s_j > 0)$, $c_j = 2^{s_j - 1}$가 되며, 이형성인 단일염기변이가 없을 때 $(s_j = 0)$, $c_j = 1$이 된다. 따라서 n개의 표본으로 구성된 자료에서 m개의 이배체형들의 도수 $n_1, n_2, \cdots, n_m$가 주어졌을 때 H개의 일배체형 비율$(p_1, p_2, \cdots, p_H)$에 대한 우도함수는 다음과 같다.

$$L(p_1, p_2, \cdots, p_H) = \frac{n!}{n_1! n_2! \cdots n_m!} \times P_1^{n_1} \times P_2^{n_2} \times \cdots \times P_m^{n_m}$$

여기서 j번째 이배체형의 비율 $P_j = \sum_{i=1}^{c_j} P(h_{ik}h_{il})$이고, $P(h_{ik}h_{il})$은 c_j개의 일배체형들의 조합들 중 i번째 조합이 일배체형 k와 일배체형 l로 구성될 확률로 임의접합(random mating)하에서 $k = l$이면 p_k^2이고, $k \neq l$이면 $2p_kp_l$이다. 각 일배체형 비율의 초기값이 설정되었다면, E-step에서 j번째 이배체형 비율은 다음과 같다.

$$P_j = \sum_{i=1}^{c_j} P(h_{ik}h_{il})$$

일배체형 k가 i번째 일배체형 조합에 $(d_{ik} = 0, 1, 2)$번 나타났다면, g번째 반복후 일배체형 k의 비율은 다음과 같이 추정된다 (M-step).

$$p_k^{(g+1)} = \frac{1}{2m} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \frac{d_{ik} P(h_{ik}h_{il})^{(g)}}{P_j^{(g)}}$$

3. 일배체형의 비율 추정

3.1. 자료

고형종양 환자 26명과 28명의 정상대조군으로부터 종양관련 사이토카인유전자 내의 1개 단일염기변이(SNP4) 와 유전자 주변의 3개의 단일염기변이(SNP1, SNP2, SNP3) 부위에 대한 유전형분석을 시행하였다. 4개의 SNP은 평균 8kb 간격으로 떨어져 있으며, 총 25 kb의 유전체 부위에 해당한다. 네 개의 단일염기변이의 대립인자는 각각 G/T, C/T, C/T, C/T이며, 이 경우 이론적으로 $2^4 = 16$개의 일배체형 종류가 가능하다. 정상군에서는 14개의 서로 다른 이배체형이 관측되었으며, 환자군에서는 8개의 서로 다른 이배체형이 측정되었다. 따라서 측정된 이배체형의 종류보다 추정하려는 일배체형의 종류가 더 많았다. 정상군과 환자군으로부터 측정된 각 단일염기변이의 이배체형의 도수는 <표 3.1>과 같다. 각 단일염기변이에서 대립유전자 및 이배체형 분포는 두 군간 차이가 없었다(Sasieni, 1997).

3.2. EM 알고리듬 적용

정상군과 환자군에서 각각 Chi square test 및 Exact test(Guo and Thomson, 1992)에 의한 Hardy-Weinberg equilibrium 성립여부 분석 결과 4개의 SNP에 대하여 각 군 모두 Hardy-Weinberg equilibrium이 성립하였다 (p-value > 0.05). 16개의 일배체형에 대한 비율을 추정하기 위하여 우도함수에 EM 알고리듬을 적용하였으며, 알고리듬은 FORTRAN 언어로 프로그래밍 되었다. 네 개의 단일염기변이에서 가능한 16개의 일배체형 각각의 초기 빈도값은 완전연관균형(complete linkage equilibrium)을 가정하여 1/16으로 하였다. 추정을 위한 알고리듬의 최대반복수는 200으로 하였으며, 수렴 기준은 0.000001로 하였다. 수렴 기준은 정상군에서 22번의 알고리듬 반복 수행 후, 환자군에서 99번의 알고리듬 반복 수행 후 만족되었다. 그 결과 각 일배체형의 비율에 대한 최대우도추정치(MLE)와 추정치의 표준오차는 <표 3.2> 및 <표 3.3> 와 같다. 표준오차는 jackknife 방법으로 추정하였다.

표 3.1: 각 단일염기변이의 이배체형의 빈도

단일염기변이	SNP1			SNP2			SNP3			SNP4		
이배체형	GG	GT	TT	CC	CT	TT	CC	CT	TT	CC	CT	TT
도수 (정상군)	8	11	9	13	13	2	2	11	15	10	9	9
도수 (환자군)	7	14	5	13	11	2	2	10	14	5	13	8
p-value[1)]	0.5586			0.8636			0.9872			0.4324		
p-value[2)]	0.4730			0.9547			0.9960			0.3039		

1) 대립유전자 분포 차이 검정 결과.
2) 이배체형 분포 차이 검정 결과.

표 3.2: EM 알고리듬에 의한 일배체형 비율의 최대우도추정치 (정상군)

일배체형	GCCC	GCCT	GCTC	GCTT	GTCC	GTCT	GTTC	GTTT
MLE	0.000000	0.000001	0.018001	0.216492	0.000000	0.193582	0.000000	0.054067
표준오차	0.000000	0.000284	0.018002	0.073785	0.000000	0.053729	0.000000	0.061122
일배체형	TCCC	TCCT	TCTC	TCTT	TTCC	TTCT	TTTC	TTTT
MLE	0.074275	0.000000	0.369659	0.018001	0.000000	0.000000	0.055922	0.000000
표준오차	0.044321	0.000000	0.079738	0.018002	0.000000	0.000000	0.053103	0.000000

표 3.3: EM 알고리듬에 의한 일배체형 비율의 최대우도추정치 (환자군)

일배체형	GCCC	GCCT	GCTC	GCTT	GTCC	GTCT	GTTC	GTTT
MLE	0.000000	0.000000	0.000000	0.250011	0.000000	0.269220	0.000000	0.019231
표준오차	0.000000	0.000000	0.000000	0.096142	0.000000	0.102598	0.000000	0.019231
일배체형	TCCC	TCCT	TCTC	TCTT	TTCC	TTCT	TTTC	TTTT
MLE	0.000000	0.000000	0.442308	0.019220	0.000000	0.000011	0.000000	0.000000
표준오차	0.000000	0.000000	0.069763	0.047535	0.000000	0.046117	0.000000	0.000000

표 3.4: 단일염기변이군간 연관불균형 검정 결과

	lnL(MLE)	lnL(EXP)[1)]	χ^2	p-value
정상군	-77.205	-113.981	73.552	< 0.0001
환자군	-51.274	-99.845	97.142	< 0.0001

1) 완전연관균형하에서의 로그우도값.

표 3.5: 단일염기변이간 연관불균형 계수(D) 및 검정 결과

단일염기변이 쌍	정상군			환자군		
	$\mid D \mid$	χ^2	p-value	$\mid D \mid$	χ^2	p-value
SNP1, SNP2	0.1013	5.4404	0.1182	0.1331	9.0332	0.0162
SNP1, SNP3	0.0644	2.3744	0.7398	0.1242	8.2090	0.0252
SNP1, SNP4	0.2317	24.1084	¡0.0001	0.2382	24.0575	< 0.0001
SNP2, SNP3	0.1123	8.5120	0.0210	0.1916	23.6280	< 0.0001
SNP2, SNP4	0.1013	5.4415	0.1182	0.1276	6.6404	0.0598
SNP3, SNP4	0.0644	2.3744	0.7398	0.1191	7.5971	0.0348

네 개의 단일염기변이에 대한 일배체형은 정상군과 환자군 모두에서 TCTC, GCTT, GTCT가 많았으며, 이들 중 일배체형 TCTC가 가장 많았고, GCTT와 GTCT의 빈도는 비슷하였다. 그러나 정상군에서는 이 세가지 일배체형이 추정된 일배체형의 비율과 완전연관균형하에서 추정된 일배체형의 비율을 비교하여 단일염기변이간 연관불균형이 존재하는지를 검정하였다. 그 결과 정상군과 환자군 모두에서 연관불균형이 있는 것으로 검정되었으며 (<표 3.4>), 두 단일염기변이간 연관불균형 계수(D)와 검정 결과는 다음과 같다 (<표 3.5>).

정상군에서는 SNP1과 SNP4간, SNP2와 SNP3간에 연관불균형이 존재하였으나, 환자군에서는 모든 단일염기변이 쌍에서 연관불균형이 존재하는 것으로 검정되었다. 이는 정상군에 비해 환자군의 일배체형 분포가 다양하지 못한 것을 뒷받침하고 있다. <표 3.2>과 <표 3.3>의 16개의 일배체형은 완전연관균형하에서 이론적으로 가능하였으나, 단일염기변이간 연관불균형에 의해 일배체형 종류들 중 일부가 빈도가 높은 소수의 일배체형으로 축소되었음을 알 수 있다.

4. 결 론

일배체형은 함께 유전되는 경향이 있는 단일염기변이군으로 단일염기변이를 그룹화 함으로써 유전체상 변이를 밝혀내기가 용이하므로 유전체 연구의 가장 중요한 기초 작업으

로 평가되고 있다. 특히 종양, 심혈관계질환 등과 같이 다수의 유전자가 복합적으로 질병에 영향을 미치는 복합성유전질환 연구에서 단일염기변이군의 일배체형에 대한 분석은 개별적인 단일염기변이에 대한 분석과 비교해 볼 때 훨씬 효과적이다. 그러나 대부분 단일염기변이군의 자료는 이배체형의 형태로 수집되기 때문에 일배체형을 알 수가 없다. 고도의 실험적 기술을 이용하는 방법과 수집된 개체의 가계를 조사하는 방법이 있지만 비용과 정확성면에서 어려움이 있다. 이때 관측된 이배체형에 대한 우도함수로부터 일배체형의 비율을 추정할 수 있다. 그러나 일배체형의 수는 대부분 관측된 이배체형의 수보다 많을 뿐만 아니라, 단일염기변이의 수가 증가함에따라 일배체형의 수는 기하급수적으로 증가하므로 일배체형 비율을 추정하는데 해석적 방법은 불가능하며 일반적인 수치적 방법보다는 EM 알고리듬의 적용이 효과적이다. 그러므로 불완전 자료에서 효과적으로 모수를 추정할 수 있는 EM 알고리듬의 적용은 일배체형 비율을 추정하는데 이배체형 유전형 분석 자료의 직접 활용을 가능케하며, 단일염기변이간 연관불균형 분석 및 일반 인구집단을 대상으로 하는 질병연관성분석에 유용할 것으로 사료된다. 본 연구에서는 고형종양 환자군과 정상군에서 네 개의 단일염기변이로 구성된 단일염기변이군의 일배체형 분석을 통하여 유전적 변이를 알아보고자 하였다. 관측된 이배체형으로부터 가능한 16개의 일배체형 비율을 효과적으로 추정하기 위하여 EM 알고리듬을 적용하였다. 이 자료를 근거로 단일염기변이 사이에 연관불균형이 존재하는지를 검정하였다. 두 군에서 모두 네 개 단일염기변이간 완전연관균형은 성립하지 않았으며, 일부 단일염기변이간에 연관불균형이 있는 것으로 검정되었다. 또한 유전자 주변주위의 일부 단일염기변이에 대해서는 정상군에서는 연관불균형이 없었으나, 환자군에서는 연관불균형이 있는 것으로 검정되어서 일부 단일염기변이와 고형종양간의 연관성에 대한 가능성을 보였다.

참고문헌

[1] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, vol. 39, 1-38.

[2] Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution*, vol. 12, 921-927.

[3] Gart, J. J. and Nam, J. A. (1984). A score test for the possible presence of recessive alleles in generalized ABO-like genetic systems. *Biometrics*, vol. 40, 887-894.

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

[5] Long, J. C., Williams, R. C., and Urbanek, M. (1995). An E-M algorithm and testing strategy for multiple-locus haplotyes. *American Journal of Human Genetics*, vol. 56, 799-810.

[6] Newton, C. R., A. Graham, L. E. Heptinstall, S. J. Powell, C. Summers, N. Kalsheker, J. C. Smith, and A. F. Markham (1989). Analysis of any point mutation in DNA: the amplification refractory mutation system (ARMS), *Nucleic Acids Research*, vol. 17, 2503-2516.

[7] Perlin, M. W., M. B. Burks, R. C. Hoop, and E. C. Hoffman (1994). Toward fully automated genotyping: allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular distrophy, *American Journal of Human Genetics*, vol. 55, 777-787.

[8] Ruano, G., K. K. Kidd, and J. C. Stephens (1990). Haplotype of mutiple polymorphisms resolved by enzymatic amplification of single DNA molecules, *Proceedings of the National Academy of Science of USA*, vol. 87, 6296-6300.

[9] Sasieni, P. D. (1997). From genotypes to genes: Doubling the sample size, *Biometrics*, vol. 53, 1253-1261.

[10] Wu, D. Y., L. Ugozzoli, B. K. Pal, and R. B. Wallace (1989). Allele-specific amplication of β-globin genomic DNA for diagnosis of sickle-cell anemia, *Proceedings of the National Academy of Science of USA*, vol. 86, 2757.

[2002년 7월 접수, 2003년 3월 채택]

Estimation of Haplotype Proportions in Single Necleotide Polymorphism Group Using EM Algorithm*

Seonwoo Kim[1] Jongwon Kim[2] Kyung-A Lee[3]

ABSTRACT

Haplotype analysis in SNP is very useful for the study of complex genetic disease due to low cost and high efficiency comparing to individual analysis of each SNP, and is functionally important in biological view. But, the gametic phase of haplotypes is usually unknown in SNP group, and it is difficult to predict haplotype proportions. In this study, haplotype proportions were estimated using EM algorithm from diploid data of SNP group in solid tumor group and normal group. From these results, linkage disequilibrium among SNPs was analyzed.

Keywords: Single nucleotide polymorphism; diploid; haplotype; EM algorithm; linkage disequilibrium

* This work was supported by grant no M10203000038-02J000002110 from the National Research Laboratory for Human and Population Genomics of the Ministry of Science & Technology.

1) Biostatistics unit, Samsung Biomedical Research Institute, Seoul, Korea
E-mail : kyuwon@samsung.co.kr
2) Department of laboratory medicine, Samsung Medical Center, School of Medicine, Sungkyunkwan university, Seoul, Korea
E-mail : jwonk@smc.samsung.co.kr
3) Department of clinical pathology, School of Medicine, Korea university, Seoul, Korea
E-mail : kal1119@hanmail.net