SNP과 Haplotype 분석의 통계적 문제점들

김호[1], 조성일[1], 서유신[2], 현순주[3], 노재정[3], 이복주[4]

요 약

Post-genome 시대를 맞이하여 인류는 전 유전체에서의 염기서열에 대한 정보를 가질 수 있게 되었다. 이러한 정보를 이용하여서 인간에게 나타나는 다양성을 설명하기 위해서 SNP(Single Nucleotide Polymorphism)의 연구가 활발히 되고 있다. 하지만 인간 체세포의 염색체는 2쌍으로 되어있기 때문에 이러한 정보가 어떠한 쌍의 조합(haplotype)으로 나타나는가를 고려하여야한다. 현재 실험적 방법으로 이를 고려하기에는 여러 가지 제약이 따르므로 통계적인 방법으로 이를 모형화하려는 노력(in silico haplotyping)이 시도되고 있다. 이 논문에서는 통계적으로 haplotype을 정하는 대표적인 알고리즘인 Clark's algorithm, E-M algorithm 등에 대한 고찰을 통하여 유전체통계학에 대한 소개를 하고자 한다.

주요용어 : Genome, SNP, Haplotype, Clark's Algorithm, EM Algorithm

1. SNP (Single Nucleotide Polymorphism)

인간 유전체(genome)는 약 30억 개의 염기서열로 이루어져있고 이 염기서열의 배열의 차이에 의해 모든 생물학적 차이가 나타난다고 알려져 있다. 그런데 이 염기서열은 사람마다 다르게 나타나고 이러한 차이를 나타내 주는 최소 단위가 SNP(Single Nucleotide Polymorphism)이다. 아래 그림은 두 염기서열을 나타내고 있는데 네모 상자가 표시된 부분에서 염기의 차이(a와 g)가 나타나서 SNP이 발견되고 있음을 보여주고 있다.

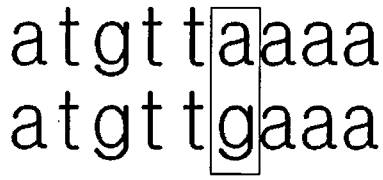

그림 1. SNP (Single Nucleotide Polymorphism)의 예

이와 같이 SNP은 가장 흔한 형태의 DNA 변이이다. 평균적으로 인간 유전체에는 1000 쌍(bp: base pair) 당 하나의 SNP이 존재한다고 알려져 있으므로 전 게놈에는 3×10^6개의 SNP이 있을 것으로 추측되고 있다. 일반적으로 특정 변이가 모집단 (population) 중 1% 이상이 나타날 때를

1) 서울대학교 보건대학원
2) 서울대학교 의과대학
3) 한국정보통신대학원대학
4) 단국대학교 전산과
이 연구는 IMT-2000 출연금기술개발지원사업(01-PJ11-PG9-01BT05-0003)의 지원으로 진행되었습니다.

polymorphism 이라 하며 1% 이하로 나타날 경우를 mutation이라 한다.

사람 개개인이 특정 질병에 대한 소인(predisposition)에 차이가 있고, 약물에 대한 반응성과 효과에 차이를 보이는 이유는 개개인 별로 게놈상에 나타나는 미묘한 차이(변이: variation) 때문이고 그들 중 가장 흔히 나타나는 변이가 SNP이다. 21세기 포스트게놈 시대에 생명과학이 당면한 가장 중요한 과제는 막대한 DNA sequence 정보를 어떻게 인류의 건강과 복지로 연결시킬 수 있느냐는 것이다. SNP은 게놈 상에 골고루 분포하고 있어 지금까지 알려진 어느 분자유전 지표들보다도 풍부하므로 복합성요인에 의해 야기되는 질병의 원인유전자를 찾아내는 데 중요한 분자지표가 된다. 이는 가계를 대상으로 하는 linkage 분석, 모집단에서의 linkage disequilibrium 연구, 환자와 대조군의 association 연구, 암환자에서의 loss of heterozygocity 연구 등에 이용된다. 또한 게놈프로젝트에서 maping 지표가 되어 자동화를 통한 해독의 속도증진에 기여하므로 SNP은 유전학 연구의 강력한 수단으로 중요한 분자지표를 제공한다고 할 수 있다. 또한 유전자의 발현조절 region이나 유전자상에 위치하는 SNP은 질병에 대한 민감성에 영향을 줄 수 있으므로 SNP과 질병의 유전적 관계가 밝혀지면 질병에 대한 민감성을 예측하고 그 질병에 대한 환경과 유전의 역할을 이해하는 데 도움을 주게 될뿐만 아니라 태아나 아직 증상이 나타나지 않은 환자에서 특정 질병에 대한 진단이 가능하며, 나아가서는 질병의 예후 및 치료, 예방에까지 이용될 수 있다. SNP은 개개인이 동일 약물에 대해 상이한 반응성-효과, 부작용, 독성-을 나타내는 지표로 사용될 수 있는데 그 원인은 다음과 같다. i) 어떤 질병이 여러 원인에 의해 (multi-allelic) 초래되는 경우 일반적으로 한 약물은 한 allele을 표적으로 하기 때문에 그 특정 allele을 가지고 있는 환자에서만 효과를 나타내게 됨. ii) 특정 약물의 대사에 관여하는 대사경로에 있는 유전자에 변이가 있는 경우 약물의 효과와 독성에 차이가 나타남. iii) 약물과 반응하는 단백질이 표적단백질과 유사한 경우 그 단백질의 기능이 마비되면서 부작용 초래. 따라서 SNP분석으로 약물의 표적을 찾아낼 수 있고, 약물에 대한 반응성을 예측할 수 있다. 많은 약물들이 preclinical trial에서 극소수의 환자에서 나타나는 부작용 때문에 임상으로까지 연결되지 않는다는 사실을 고려하면 SNP 분석을 통해 중요한 약물-개발될 때까지 막대한 비용이 소요된-을 구제하고, 부작용으로 사망할 환자를 구제할 수 있다.

이러한 SNP 분석의 중요성과 응용성 때문에, 1997년부터 산업계와 학계에서 SNP 발굴을 위한 대규모의 노력을 진행하고 있다. 대표적인 예로 1997년에 DNA microchips 생산회사인 Affymatrix와 MIT가 공동으로 $40,000,000을 투자하여 2000개의 SNP를 찾는 것을 목표로 시작한 이후 Abbot과 Genset이 60,000개의 SNP를 목표로 $43,000,000을 출자하였고, 1998년에 Incyte가 $200,000,000을 NHGRI가 $30,000,000을, 1999년에는 일본정부가 2년 동안 150,000개의 SNP를 목표로 $50,000,000을, 비영리 SNP 콘소시움인 TSC가 300,000개의 SNP를 목표로 $45,000,000을 투자하였다. 2002년 말까지 300,000개의 SNP이 발굴될 예정이다. 질병의 원인 유전자를 발굴하기 위해서는 약 3000 bp 당 한 개의 SNP 즉 1,000,000 SNPs가 필요할 것으로 추정된다. 하지만 2000년 3월 현재 50,000 SNP만이 보고되었다. 빠른 기술의 진보 때문에 예상보다 빠른 속도로 SNP 발굴이 진행되겠지만, 목표에 도달하기 위해서는 SNP 발굴을 위한 계속적인 노력이 필요하다. 뿐만 아니라 특정 SNP의 유용성은 상당한 인종적 차이를 나타내어, 지금까지 알려진 SNP 중 약 3분의 1만이 인종적 차이를 나타내지 않는 것으로 밝혀지고 있다. (Stephens et al. 2002a) 이는 특정 인종그룹에 의미 있는 SNP은 그 수가 훨씬 많으리라는 것을 시사한다. 따라서 한국인 특유의 유전적 배경을 바탕으로 한국인에 유용한 SNP를 발굴하는 것이 절실히 요구된다. 표1은 우리나라의 대표적인 SNP연구인 21C Frontier연구의 연구목표를 나타내고 있다.

표1. 21C Frontier 사업단의 SNP 발굴 목표

구 분	1단계(2000-2002)	2단계(2003-2005))
1. 한국인 SNP발굴 및 지도작성	5,000	5,000
2. 위암.간암 유전자 cSNP발굴	130	250
3. 주요질환 유전자 cSNP발굴	110	130
발굴목표(개)	5,240(개)	5,380(개)
배정예산(억원)	10(억)	13(억)

2. Haplotype

일정 부위에 polymorphism(SNP)이 존재하게되면 한 쌍의 염색체에 두 가지 다른 형태의 염색체가 존재할 수 있으므로 이러한 염색체의 상태를 allele(대립유전자)이라고 한다. 이때 각각의 allele이 나타나는 빈도가 각각의 SNP에 따라서 다르며 이는 또한 인종에 따라 다른 빈도로 나타나므로 특정 SNP이 특정 인구에서 나타나는 빈도 (frequency), 즉 allele frequency는 각각의 SNP 마다 또한 인종마다 특이한 분포를 보인다. 발굴된 SNP이 중요한 표지인자로서 유용하게 활용되려면 인구(population)내에 일정 빈도 이상으로 다형성이 나타나야 한다. SNP의 유용성을 확인(validation)하는 가장 중요한 과정중의 하나가 각각의 SNP들에 대한 allele frequency를 밝히는 일이다. 최소한 10%이상 정도의 빈도로 나타나는 SNP들만이 다음 단계인 특정 질병과의 연관관계를 밝히는데 유용하게 활용될 수 있을 것으로 예상되고 있다. 최근 단일 SNP이 의학적으로 유용한 표현형 연관분석에 무용하며 여러 SNP들이 한 염색체상에 연결되어 있는 SNP의 조합인 Haplotype이 더 많은 정보를 포함하는 것으로 알려지고 있다. (Daly et al. 2001). 그림2는 allele frequencies 만으로는 haplotype을 결정할 수 없음을 보여주고 있다. Haplotype structure의 파악은 질병과 연관되어 있는 유전자의 위치를 연구하기 위한 정확한 방법을 제공한다. 질병에 영향을 주는 유전자들은 haplotype block 단위로 변이를 가지기 때문에, 개별 SNP을 이용한 기존의 association study 방법은 통계적 검정력이 약하며 분명한 위치 파악이 어려운 것에 비해, haplotype block approach는 유전자 내의 모든 주요 변이에 대해서 질병과의 연관성을 빠짐없이 평가하는 것이 가능하다. 또한 유용한 SNP 발굴을 위한 단일 SNP 분석은 수천명 단위의 genotyping이 필요하지만 haploytpe 분석은 수백명 단위로 필요한다고 알려져 있다. 이러한 이유들로 NIH를 중심으로 post genome research의 방향으로 설정하여 막대한 연구 인력이 동원되어 진행되고 있다. Orchid Bioscience, Genaissance, Pearlgen 등의 bioventure를 중심으로 haplotype mapping 경쟁이 치열한 상태이다. 이러한 Haplotypedml 분석으로는 실험적으로 haplotype을 결정하는 moldecular haplotype과 통계적 방법으로 결정하는 in silico haplotype이 있다. Molecular haplotyping은 실험적으로 같은 염색체상의 SNP 조합을 결정하므로 정확하게 haploytpe을 분석하고 빈도가 낮은 haplotype의 발굴이 가능하나 비용 및 시간이 많이 들고 기술적인 어려움이 있다. In silico haplotying (haplotype reconstruction)은 genotyping 결과에 근거하여 알고리즘을 통하여 통계적으로 예측하므로 시간 및 비용이 절감되지만 정확성이 떨어질 수 있고 많은 수의 샘플이 분석되어야 의미 있는 결과를 도출할 수 있고 빈도가 낮은 haplotype의 발굴이 어렵다는 단점이 있다. 표2는 이 두 가지 방법을 요약하고 있다.

표2. Haplotying의 방법들

Molecular Haplotyping
Molecular Cloning: Clasper and TAR 기술 (150kb 이상의 DNA) Single Molecule Dilution (SMD) Allele Specific PCR Heteroduplex Analysis Mismatch Detection Allele-Specific Oligonucleotide Hybridization
In-silico Haplotyping
Clark's Algorithm (Clark, 1990): HAPINFERX program EM (Expectation-Maximization) Algorithm (Hawley & Kidd, 1995): HAPLO program Pseudo-Baysian Algorithm (Stephens et al., 2001b): PHASE program

possible alleles	possible haplotypes
A/a	ABC
B/b	ABc
C/c	AbC
	Abc
	aBC
	aBc
	abC
	abc

그림2. 세 bialleic loci A, B, C들이 만들 수 있는 haplotype의 조합들

3. In-Silico Haplotype Determination Methods

이 절에서는 In-silico haplotype의 대표적 방법인 E-M algorithm과 Clark's algorithm을 살펴보기로 하겠다. 우선 n개의 diploid individuals이 있다고 하자. $G=(G_1,\cdots,G_n)$은 (알려져 있는) genotype이라고 하고 $H=(H_1,\cdots,H_n)$은 (미지의) haplotype pairs라고 하자. $F=(F_1,\cdots,F_M)$은 (미지의) population haplotype frequencies이고 $f=(f_1,\cdots,f_M)$은 sample haplotype frequencies이다. . E-M 알고리즘(Excoffier and Slatkin, 1995)은 다음의 likelihood를 최대화하는 F를 찾는 것이다.

$$L(F)=\Pr(G|F)=\prod_{i=1}^{n}\Pr(G_i|F)$$

여기서 $\Pr(G_i|F) = \prod_{(h_1,h_2)\in H_i} F_{h_1}F_{h_2}$이고 H_i는 multilocus genotype G_i에 해당하는 haplotype pairs의 집합이다. 이 likehood는 Hardy-Weinberg equilibrium의 가정 하에서 sample genotype의 확률을 population haplotype frequencies의 함수로 표현한 것이다.

Clarks' algorithm (Clark 1990)은 표본에서 관찰된 haplotype의 총 개수를 최소화하려는 것으로 일종의 parsimony approach이다. 이 알고리즘은 표본에서 확정적으로 결정되는 haplotype들을 열거하면서 시작한다. 알려진 확정적 haplotype을 가지고 미지의 haplotype과 결합한 관찰된 haplotype을 구축하고 이러한 과정을 모든 자료들에게 haplotype이 열거될 때까지 반복한다.
이 두 가지 이외에도 Stephens등 (2000b)이 제안한 pseudo-Gibbs sampler (PGS), Niu 등 (2002)이 제안한 Bayesian Haplotype Inference 등의 방법이 있다.

4. Discussion

유전체 연구 등 현대의 생물학에서 SNP의 연구는 가장 중요하고 실용적 가치가 있는 분야로 인정되고 있다. (Horvath & Baur 2000) 이러한 SNP의 연구에서 haplotype을 고려해야 한다는 것이 연구의 복잡성을 더해주고 있고 실험적 방법으로 이를 고려하기에는 여러 가지 제약이 따르므로 통계적인 방법으로 이를 모형화하려는 여러 가지 노력이 있어왔다. 하지만 다른 bioinformatics분야와 마찬가지로 어떠한 모형이 우수한지에 대해서는 아직도 많은 연구가 진행되고 있다. SNP 정보는 인종간에 상당한 변이가 있다고 알려져 있으므로 우리나라 인구를 대상으로 한 연구가 많이 진행되고 있으며 앞으로는 이러한 정보 자체가 국가경쟁의 큰 부분이 될 것은 자명한 일이다. 실험실에서 SNP의 기초 자료를 생산하는 일 못지 않게 이러한 자료에서 유전정보학적 의미를 뽑아내는 일이 중요하고 이를 위하여서 많은 통계학자들이 이러한 연구에 참여해야한다고 판단된다.

참고문헌

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111-112.

Daly et al (2001) High-resolution haplotype structure in the human genome. Nature genetics 29:229-232.

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular hyplotype frequencies in a diploid population. Mol Bio Evol 12: 921-927.

Hawley ME, Kidd KK (1995) HAPLO: a program using the Em algorithm to estimate the frequencies of multi-site haplotypes. J Hered 80: 409-411.

Horvath S, Baur MP (2000) Future directions of research in statistical genetics, Statistics in Medicine 19: 3337-3343.

Stephens J.C. et al (2001a). Haplotype variation and linkage disequilibrium in 313 human genes, Science 298: 489-493.

Stephen M, Smith NJ, Donnelly P (200b) A new statistical method for haplotype reconstruction from population data, Am J Hum Genet 68: 978-989.

Niu T, Qin ZS, Xu X, Liu JS (2002) Bayesin haplotye inference for multiple linked single-nucleotide polymorphisms, Am. J. Hum. Genet. 70: 000-000.