결측치가 존재하는 유전형 자료에서의 연관불균형과 일배체형을 사용한 결측치 대치 방법

(A New Method for Imputation of Missing Genotype using Linkage Disequilibrium and Haplotype Information)

박 윤 주 † 김 영 진 † 박 정 선 †
(Yun-Ju Park) (Young-Jin Kim) (Jung-Sun Park)
김 규 찬 † 고 인 송 †† 정 호 열 †††
(Kuchan, Kimm) (InSong Koh) (Ho-Youl Jung)

요 약 본 논문에서는 단일염기변이(SNP: Single Nucleotide Polymorphism)와 같은 유전형(genotype)자료에서 결측치가 발생하였을 경우 유전형 자료의 특이성을 고려해 자료 원래의 정보손실을 최소화하는 대치법인 연관불균형 기반의 대치법(linkage disequilibrium- based imputation)과 일배체형 기반의 대치법(haplotype-based imputation)을 제시한다. 이러한 결측치 대치는 실험상에서 발생하는 결측치에 의한 중요한 정보의 손실을 최소화 한다는 점에서 필요한 방법이다. 일반적으로 그동안 생물학 자료의 결측치 대치는 대부분 주형질 대치법(major allele imputation)이 활용되어왔는데 유전형 자료에서의 이 방법의 사용은 자료의 특이성으로 인하여 결측치에 대한 높은 오차율(error rate)을 보임으로서 자료의 신뢰성을 떨어뜨릴 수 있다. 본 논문에서는 유전형 자료인 단일염기변이 자료의 시뮬레이션을 통하여 기존의 주형질 대치법과 논문에서 제안된 연관불균형 기반의 대치법과 일배체형 기반의 대치법을 비교하고 그 결과를 보여 준다.

키워드 : 생물정보학, 단일염기다형성, 이배체형 대치법

Abstract In this paper, we propose a new missing imputation method for minimizing loss of information- linkage disequilibrium-based and haplotype-based imputation method, which estimate missing values of the data based on the specificity of Single Nucleotide Polymorphism(SNP) genotype data. Method for imputing data is needed to minimize the loss of information caused by experimental missing data. In general, missing imputation of biological data has used major allele imputation method, but this approach is not optimal. This method has high error rates of missing values estimation since the characteristics of the genotype data are not considered not take into consideration the specific structure of the data. In this paper, we show the results of the comparative evaluation of our model methods and major imputation method for the estimation of missing values.

Key words : bioinformatics, single nucleotide polymorphism, genotype imputation

1. 서 론

인종간 또는 개개인이 특정 질병에 대한 차이가 있고, 약물에 대한 반응성과 효과에 차이를 보이는 이유는 유전체에 나타나는 변이로서 설명되고 있다. 이 중 가장 흔히 나타나는 변이인 단일염기변이(single nucleotide polymorphism: SNP)는 인간 질병에 관여하는 것으로 보이는 다형(polymorphism)의 주를 이루는 변형(variant)으로, 정보화 가능 부분(coding)과, 비정보화 부분(noncoding region)에서 약 천개의 염기 당 1개의 빈도

† 비 회 원 : 국립보건연구원 유전체연구부
oct1001@ yahoo.co.kr
in time@hanmail.net
jspark518@ngri.re.kr
k2kimm2@nih.go.kr
†† 비 회 원 : 과학기술혁신본부 기술혁신평가국 보건연구관
insong@nih.go.kr
††† 정 회 원 : 한국전자통신연구원 바이오정보연구팀 연구원
hyj@etri.re.kr
논문접수 : 2004년 7월 29일
심사완료 : 2004년 11월 23일

로 매우 빈번하게 관찰 된다[1].

단일염기변이는 3가지의 이유에서 그 의미를 가지는데, 첫째 특정한 기능을 가지는 유전자에서 발견이 되는 단일염기변이의 경우 그 다형성의 변화가 단백질의 형태나 발현(expression)에 영향을 주어 병을 유발하거나, 또는 다른 형태로 표현형(phenotype)의 차이를 보여준다. 둘째, 단일염기변이는 유전적 표지 혹은 표식유전자(genetic marker)의 역할로 관심을 가지고 있는데, 표현형에 관련된 유전적 차이를 추측할 수 있는 역할을 한다. 마지막으로, 돌연변이 비율과 관련된 연구와 진화역사(evolutionary history)의 연구에 유용하게 사용된다[2].

모든 자료에서 결측치(missing values)의 문제는 빈번하게 발생하며, 이와 같은 결측치가 있는 자료를 분석할 경우 일반적으로 결측치가 있는 개체들을 제거하거나 또는 주어진 다른 정보를 이용하여 결측치를 대치하는 방법이 일반적이다. 이렇게 주어진 정보를 이용하여 결측치를 대치하는 방법을 대치법(imputation)이라고 한다.

이러한 현상은 임상자료를 다루는 의학, 보건학 분야뿐 아니라 생물학 분야의 모든 자료에도 공통적으로 나타나고 있는 실정이다[5-8]. 유전형 자료인 단일염기변이 데이타의 경우에도 이와 같은 결측치가 존재하는데 이러한 결측치의 비율이 높을수록 자료의 신뢰성이 떨어져 잘못된 결과를 도출하게 될 가능성이 커진다.

본 논문의 실험에서 사용하게 될 자료인 International HapMap Project의 1~22번까지 염색체(chromosome)의 유전형 자료에서 결측치률을 살펴보면 표 1과 표 2와 같다. HapMap의 유전형자료는 그림 1과 같이 단일염기변이 데이타로 각 염색체별 90명이 있고 90명중 5명은 반복실험 되어 있다. NA12003.dup같이 개인번호 뒤의 .dup 라고 표시되어있는 부분은 반복 실험을 나타낸다. 데이타의 행 부분은 개인번호이고 열 부분은 단일염기변이 사이트로 구성되어 있다[14]. 위 그림에서 AG, GG와 같은 대립유전자 쌍은 단일염기변이를 의미하는데, 단일염기변이 rs1044085사이트에서 NA11840 개인번호의 단일염기변이 AG같이 두 대립유전자가 다른 것을 이형접합자(heterozygote)라고 부르고 NA11881의 GG같이 두 대립유전자가 G, G로 같은 유전자를 동형접합자(homozygote)로 부른다. 사람의 염색체는 부모에게서 각 각 받은 한 쌍의 이배체형 서열(diploid sequence)로 구성되어 있는데, 단일염기변이가 발생하는 부위의 이형접합자 같은 경우는 부모의 대립유전자(allele)가 서로 다른 경우, 즉 예를 들면 부 쪽의 대립유전자쌍이 AA이고 모 쪽의 대립유전자쌍이 GG이거나 반대로 모 쪽의 대립유전자 쌍이 AA이고 부 쪽이 GG이면 자녀는 AG 대립유전자 쌍을 가지게 된다. 만약 부 쪽과 모 쪽의 대립유전자 쌍이 같다면 자녀는 AA나 GG와 같은 동형의 대립유전자를 가지게 된다.

표 2는 각 염색체별 5명의 반복 실험한 개인에서 단일염기변이 데이타의 결측치률을 보여주고 있다. 여기서

표 1 HapMap Project 1~22번 염색체의 유전형 자료의 결측치률

데이타의 총합(개)	33,151,320
데이타의 결측치 개수(개)	382,428
전체 데이타의 결측치률(%)	1.15

표 2 HapMap Project 1~22번 염색체에서의 반복 실험한 결측치률

결측치 -> 결측치 (개)	7,344
결측치 -> 비 결측치(개)	24,434
반복된 실험에서의 결측치률(%)	23.11
반복된 실험에서의 비결측치률(%)	76.89

rs#	rs1044085	rs13750	rs1049536	rs1061541	rs1057457	rs1547411	rs135021	rs1984388	rs4700
NA11840	AG	CC	CT	TT	AG	TT	AA	TT	CT
NA11881	GG	CT	CC	TT	AG	TT	TT	TT	CT
NA11882	GG	CT	CT	CT	AG	CT	AT	TT	CC
NA11992	GG	CT	CC	TT	GG	TT	TT	AT	CT
NA11993	AG	CC	CT	TT	AA	TT	AT	TT	CT
NA11993.dup	AG	CC	CT	TT	AA	TT	AT	TT	CT
NA11994	GG	CT	CC	TT	AG	TT	TT	AT	CT
NA11995	AG	CT	CC	TT	GG	TT	AA	TT	CC
NA12003	GG	CC	CT	TT	AG	TT	AT	TT	TT
NA12003.dup	GG	CC	CT	TT	AG	TT	AT	TT	TT
NA12004	GG	CC	CC	CT	AA	CT	AT	TT	CT
NA12005	GG	CT	CC	TT	AA	TT	AT	AT	CC
NA12006	GG	CC	CC	TT	AG	TT	AT	TT	CC
NA12043	GG	CC	CT	CT	GG	TT	AT	TT	CT
NA12044	GG	CC	CC	CT	AG	TT	TT	TT	CT
NA12056	GG	CT	TT	CC	GG	TT	TT	TT	TT
NA12057	GG	CT	CT	TT	AG	CT	AT	TT	CT
NA12144	GG	CT	TT	CT	AG	TT	TT	TT	TT
NA12145	GG	CT	CC	CT	AG	TT	AA	TT	CT
NA12146	GG	CC	CC	TT	GG	TT	TT	AT	CT
NA12154	GG	CC	CC	TT	AA	TT	AT	TT	TT

그림 1 International HapMap Project의 유전형 자료

보듯이 반복실험에 의하여 결측치가 비결측치로 나올 확률은 76.89%로 이 수치는 실험상의 실수나 소프트웨어의 결점 같은 기타 요인에 의한 자료의 결측치가 발생할 가능성이 높다는 것을 보여주고 있다. 반면에 결측치가 결측치로 나오는 것은 23.11%로 상당히 높은 수치인데, 이는 위와 같은 문제에 의해서라기보다 염색체 고유의 특성에 의해서 자료가 추출되지 않는 경우가 상당히 많다는 것을 보여주는 결과로 이러한 경우 반복실험을 하여도 비 결측치가 나올 확률이 낮아진다. 일반적으로 이렇게 반복하여도 결측치가 생기는 경우 대부분의 해결방법으로 결측치가 발생한 단일염기변이 사이트를 제외하고 사용하였는데, 만약 이러할 경우 제외한 사이트가 질병에 관련 되어있거나 연구에 필요한 정보를 담고 있는 중요한 단일염기변이 사이트라면 심각한 잘못된 분석결과를 초래할 수 있을 것이다.

이러한 문제 때문에 대치법을 사용하는데, 대치법은 주어진 자료의 정보를 고려하여 추정하는 경우가 결측치가 발생한 개체를 제외하는 경우와 비교하여 보았을 때 원 자료의 정보의 손실이 적기 때문에 그 동안 널리 사용되었고 또한 폭넓게 개발되어 왔다[5-7].

2. 단일염기변이자료의 대치법 문제

기존에 일반적으로 생물학적 자료의 결측치 대치에 사용되는 몇 가지 방법들이 있는데, 첫 번째로 결측치를 제외한 나머지 자료의 평균을 구하여 결측치에 평균값을 대치시키는 평균대치법이 있고, 두 번째로 통계학적인 개념을 도입한 K Nearest Neighbors(KNN)대치법, SVD(Singular Value Decomposition)대치법, EM알고리즘(Expectation Maximization Algorithm), Bayesian principal component analysis(BPCA) 대치법 등이 있다[9].

이런 생물학자료 중 단일염기변이와 같은 범주형 자료인 유전형자료의 대치법으로는 결측치를 제외한 자료에서 확률적으로 가장 많이 나타나는 대립유전자를 대치시키는 주형질 대치법이 주로 사용되고 있다. 그러나 이 주형질 대치법은 알고리즘 이 간단하고 계산속도가 빠르다는 장점이 있으나 유전형 자료의 특이성을 반영하지 못해 오차율(error rate)이 높다는 가장 큰 단점이 있다. 이러한 유전형 자료는 다른 생물학적 자료와는 달리 그림 2와 같이 연관성이 높은 단일염기변이 사이트들 끼리 블록을 구성하여 전체 자료를 구성하고, 이러한 높은 연관성을 보이는 구간에서 제한적인 일배체형의 다양성(haplotype diversity)을 보여주고 있기 때문에, 연관성 블록을 고려하지 않은 전체 자료를 대상으로 사용되고 있는 주형질 대치법을 사용하게 된다면 연관성 블록에서 나타나는 공통된 일배체형(common haplotype)에 심각한 영향을 주게 되어 잘못된 결과를 도출할 가능성이 높아진다.

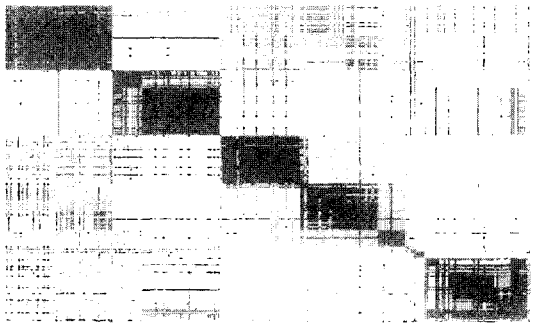

그림 2 International HapMap Project의 22번 염색체의 rs204970~rs713705 단일염기변이 사이트끼리의 연관성 블록: 행과 열은 단일염기변이 사이트를 나타내고 각 칸은 단일염기 변이 사이트들끼리의 연관성 수치인 |D'|값을 나타낸다. 칸의 색이 짙어질 수 록 연관성 수치가 높다(그림 2에서 검정색 부분이 |D'|>0.8 이다). 사각형으로 표시된 부분이 하나의 연관성 블록을 이루는 것이다.

본 논문에서는 이러한 단점을 보완하여 단일염기변이 같은 유전형 자료의 특성을 고려하여 가능한 원 자료의 손실을 최소화하면서 결측치를 대치하는 새로운 방법인 일배체형 기반의 대치법과 연관불균형 기반의 대치법을 제시한다.

3. 대치법(Imputation)

단일염기변이는 두 개의 대립유전자(bi-allele)이상으로 구성되어 있지만 일반적으로 두 개의 대립유전자로 구성되어있는 경우가 가장 흔하다[10]. 본 논문에서 쓰이는 단일염기변이 자료는 두 개의 대립유전자를 갖는다고 가정하였다.

현재의 유전형확정법(genotyping)이나 서열분석법(sequencing method)은 이배체형 서열인 염색체 단위로 분석을 하기 때문에 일배체형의 위상정보를 제공하지 못한다. 비록 체세포 융합 같은 기술을 통해서 가능하다고는 하지만, 실험적인 오류의 가능성은 남아있다. 이 때문에 대부분의 단일염기변이 데이타는 위상구분이 되어있지 않다[3]. 예를 들면 위의 그림 1의 설명에서 보듯이 이러한 대립유전자의 쌍으로 구성되어있는 단일염기변이는 AG같은 이형접합자인 경우 부모의 자료가 있는 가계도 자료(pedigree data)가 존재하지 않으면 각 대립 유전자가 두 부모 중 어느 쪽에서 유전되었는지에

대한 위상(phase)을 알 수 없다. 이 때문에 이형접합자가 많아질 경우 유추 가능한 일배체형의 조합이 기하급수적으로 늘어나는데, 이러한 이유 때문에 단일염기변이 데이타 같은 유전형 자료에서 일배체형(haplotype)을 추정하여 자료를 재구성(reconstruction)하여야 한다. 본 논문에서는 그림 3에서 보여주는 것처럼 EM 알고리즘을 사용하여 단일염기변이 데이타에서 일배체형을 추정하였다[11,13].

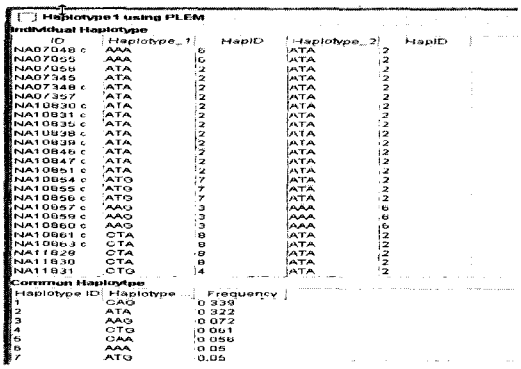

그림 3 EM 알고리즘을 사용하여 추정한 단일염기변이의 일배체형

3.1 일배체형 기반의 대치법

일배체형이란 밀접하게 연관된 대립유전자의 집합을 말한다. 일배체형은 이배체형 서열에서 이형접합자의 서로 다른 대립유전자의 관계를 말하며 이형접합자의 가능한 조합에서 각 개체가 가질 수 있는 공통된 일배체형(common haplotype)을 결정한다. 본 논문에서 제시하는 첫 번째 결측치 대치 방법인 일배체형 기반의 대치법은 연관성이 높은 블록에서는 제한된 공통 일배체형을 가진다는 것에서 고안한 방법으로, 대치 알고리즘은 아래와 같다.

결측치를 가지고 있는 개체의 경우는 결측치를 제외한 자료에서 얻을 수 있는 일배체형 쌍을 도출하고, 도출된 공통 일배체형들 중 결측치가 있는 개체에서 도출한 일배체형과 비교하여 확률값이 높은 일배체형을 결측치에 대치시킨다.

표 3 단일염기변이의 이배체형 서열과 추정할 수 있는 일배체형: N/N은 결측치이다.

이배체형 서열			일배체형쌍
A/T	C/G	A/T	ACA/TGT
A/T	G/G	T/T	AGT/TGT
A/A	C/C	T/T	ACT/ACT
A/T	N/N	A/T	ANA/TNT
T/T	G/G	T/T	TGT/TGT

위의 자료에서는 ACT, TGT, AGT, ACA가 공통된 일배체형임을 알 수 있다. 결측치가 존재하는 개체는 ANA/TNT의 일배체형을 가진다. 대치될 수 있는 일배체형으로는 ANA의 경우 ACA가 되고, TNT의 경우는 TGT가 된다. 따라서 C/G가 대치된다.

만약 결측치의 일배체형이 ANT/TNT 으로 추정된 것이 있었다면 위표에서 대치될 수 있는 일배체형으로는 ANT의 경우 ACT와 AGT가 되고, TNT의 경우는 TGT가 된다. ACT의 경우 자료에서 보이는 8개의 일배체형 중에서 3번 나타나지므로 3/8의 확률을 보여주고 있고, AGT의 경우는 1/8의 확률을 보여주고 있다. 따라서 확률값이 높은 ACT가 가장 적합한 것으로 꼽힌다. 결과적으로 결측치는 C/G로 대치된다.

3.2 연관불균형(linkage disequilibrium) 기반의 대치법

결측치 자료를 대치하는 두 번째 방법으로 연관불균형 기반의 대치법이 있다. 본 논문에서 언급하는 연관성 불균형이란 유전형 자료에서 단일염기변이 사이트간의 상관성을 나타내는 수치로서 여러 가지 방법들이 이미 알려져 있다[18]. 이 수치가 높을수록 단일염기 변이 사이트간의 연관성이 높은 것으로, 이 경우 사이트들의 대립유전자들은 서로 연관되어서 항상 같은 비율로 나타나게 된다. 예를 들어 만약 그림 1에서 첫 번째 단일염기 변이 사이트 rs1044085와 두 번째 rs13750 사이트가 완전한 연관성을 보인다고 가정한다면 첫 번째 사이트에 AG와 두 번째 사이트 CC는 서로 연관되어 같은 비율로 나타나게 된다.

이러한 연관불균형 수치에 따라 연관성이 높은 단일염기변이 사이트끼리 하나의 연관성 블록을 이룰 수 있는데 이 블록 안에서 결측치를 대치할 수 있다. 즉, 이러한 유전형 자료는 여러 개의 낮은 연관불균형 지수를 가지는 재조합 열점(recombination hotspot)이라고 불리는 분화된 조각에 의해 나누어지는 높은 연관불균형 지수를 가지는 부분들로 이루어져 있다. 이 높은 연관불균형 지수를 가지는 부분에서는 제한된 일배체형의 다양성을 보이고 있으며, 공통된 일배체형을 가지고 블록에 나타나는 대부분의 유전자를 설명할 수 있다. 이러한 제한된 일배체형 다양성을 가지는 높은 연관불균형 지수를 보이는 구간을 일배체형 블록(haplotype block)이라고 하는데[4], 이러한 블록의 연관성 불균형 정보를 이용하여 결측치를 대치하면 결측치 대치시 오차율을 낮출 수 있다.

연관불균형 측정 알고리즘은 현재까지 알려진 여러 가지 방법들이 있는데 가장 보편적으로 Lewontin이 제안한 연관성 값인 D'이 널리 쓰이고 있다[12]. 이 D'값은 단일염기변이 사이트간의 연관성정도를 나타내는 값

으로 0에 가까울수록 두 단일염기 사이트가 서로 독립되어 있다고 볼 수 있다. Lewontin의 D′은 아래와 같이 정의된다.

Allele	B	b	
A	n_{AB}	n_{Ab}	n_{A+}
a	n_{aB}	n_{ab}	n_{a+}
	n_{B+}	n_{b+}	n

$$D = p_{AB} - p_{A+}p_{B+}$$

$$D' = \begin{cases} \dfrac{D}{\min(p_{A+}p_{b+} : p_{a+}p_{B+})} & D>0 \\ \dfrac{D}{\min(p_{A+}p_{B+} : p_{a+}p_{b+})} & D<0 \end{cases}$$

위의 왼쪽 표에서 n_{AB}는 첫 번째 대립유전자가 A이고 두 번째 대립유전자가 B인 일배체형 자료의 개수이고, n_{A+}는 A를 첫 번째 대립유전자로 가지는 일배체형의 개수, n_{B+}은 B를 첫 번째 대립유전자로 가지는 일배체형의 개수, 그리고 n은 총 일배체형의 개수이다. 오른쪽 D 식에서 표시된 p_{AB}는 n_{AB}를 총 수 n로 나눈 확률값으로 $p_{AB} = n_{AB} / n$로 표현되고, $p_{A+} = n_{A+} / n$로 표현한다. 본 논문에서는 위에서 언급한 Lewontin의 방법을 기반으로 하여 유전형 자료의 결측치 대치를 하였다.

앞 절에서 언급하였던 일배체형 기반의 대치법과 마찬가지로 단일염기변이 데이타에서 결측치가 발생한 위치의 행을 제외하고 결측치가 없는 나머지 데이타의 일배체형을 도출해낸 후 이 일배체형 자료로 아래와 같이 {첫 번째 대립유전자, 두 번째 대립유전자}, {두 번째 대립유전자, 세 번째 대립유전자}쌍의 대립유전자 빈도표(frequency table)를 작성한다. 예를 들어 보면 표 3에서 결측치가 발생한 4번째 행의 이배체형 서열인 A/T N/N A/T를 제외한 후 결측치가 없는 나머지 이배체형 서열을 가지고 일배체형을 추정하여 아래와 같이 대립유전자 빈도표를 작성한다.

first allele \ second allele	B	b	
A	n_{AB}	n_{Ab}	n_{A+}
a	n_{aB}	n_{ab}	n_{a+}
	n_{B+}	n_{b+}	n

second allele \ third allele	C	c	
B	n_{BC}	n_{Bc}	n_{B+}
b	n_{bC}	n_{bc}	n_{b+}
	n_{C+}	n_{c+}	n

위 대립유전자 빈도표를 바탕으로 대립유전자 확률표를 아래와 같이 구성할 수 있다.

first allele \ second allele	B	b	
A	p_{AB}	p_{Ab}	p_{A+}
a	n_{aB}	p_{ab}	p_{a+}
	p_{B+}	p_{b+}	1

second allele \ third allele	C	c	
B	p_{BC}	p_{Bc}	p_{B+}
b	p_{bC}	p_{bc}	p_{b+}
	p_{c+}	p_{c+}	1

위의 {첫 번째 대립유전자, 두 번째 대립유전자}, {두 번째 대립유전자, 세 번째 대립유전자}의 확률표에서 각각 Lewontin의 D′값을 계산한다. 원 식에서 D′값이 1에서 1 값의 범위를 취하기 때문에 본 논문에서는 부호를 배제하고 단일염기변이 사이트끼리의 연관불균형 정도만을 알아보기 위하여 |D′| 값을 계산한다.

위 표에서 계산된 |D′|과 대립유전자 확률값을 가지고 단일염기변이 사이트에서 추정된 일배체형 자료로 사이트간의 연관성 값을 계산할 수 있다. 예를 들어 세 개의 단일염기변이 사이트에서 ABC라는 일배체형이 추정되었다면 세 사이트의 연관성 값은 ($p_{AB} \times |D'|_1$) + ($p_{BC} \times |D'|_2$)으로 계산된다. 여기에서 $|D'|_1$은 위 대립유전자 확률표의 왼쪽 표에서 계산된 |D′|값이고 $|D'|_2$는 오른쪽 표에서 계산된 값이다.

마지막으로 단일염기변이 데이타에서 결측치가 발생한 행의 일배체형 자료를 추정해 본 후 앞에서 계산 된 연관성 값이 가장 큰 일배체형의 부분을 결측치 부위에 대치시키게 된다. 표 3을 연관불균형 기반의 대치법을 이용하여 결측치를 대치시켜보면 아래와 같이 계산된다.

first allele \ second allele	G	C	
A	1	3	4
T	4	0	4
	5	3	8

second allele \ third allele	T	A	
G	5	0	5
C	2	1	3
	7	1	8

우선 표 3에서 결측치를 제외한 나머지 8개의 일배체형 으로 대립유전자 빈도표 작성해보면 위와 같고, 이를 바탕으로 대립유전자 확률표를 작성해본다.

first allele \ second allele	G	C	
A	1/8= 0.125	0.375	0.5
T	0.5	0	0.5
	0.625	0.375	1

second allele \ third allele	T	A	
G	0.625	0	0.625
C	0.25	0.125	0.375
	0.875	0.125	1

위 결과에서 $|D'|_1$과 $|D'|_2$을 계산해보면 $|D'|_1$=|-0.1875/0.1875|=1이고 $|D'|_2$=|0.078125/0.078125|=1이다. 마지막으로 공통된 일배체형인 ACT, TGT, AGT, ACA의 연관성 값을 계산해보면, ACA= (0.375×1)+(0.125×1)=0.5, 마찬가지 방법으로 AGT=0.75, ACT=0.625, 그리고 TGT=1.125이다. 표 3의 결측치에 대치될 수 있는 일배체형으로는 ANA의 경우 ACA가 되고, TNT의 경우는 TGT가 된다. 따라서 C/G가 대치된다. 만약 결측치가 ANT로 추정되었다면 일배체형 기반의 방법과 마찬가지로 AGT와 ACT중 연관성 값이 큰 AGT의 G로 대치된다.

4. 실험 결과

실험에서 사용한 데이타는 크게 2가지로 구분할 수 있다.

1) 가상 데이타 : 아래와 같은 순서로 가상데이타를 생성하였다.
   가) 세 개의 단일염기변이 사이트에 대한 유전형자료를 International HapMap Project 데이타의 명수와 같이 90명 랜덤 하게 생성 하였다.
   나) 두 개의 쌍 {첫 번째 사이트, 두 번째 사이트}, {두 번째 사이트, 세 번째 사이트}에 대한 D'의 신뢰구간을 측정 하였다.
   다) 나)에서구한 신뢰구간이 원하는 신뢰구간을 만족하지 않으면, 가)의 단일 염기변이 사이트의 유전형자료를 원하는 신뢰구간을 만족할 때까지 실험자가 직접 수정하였다.

위와 같은 방법으로 아래와 같이 3가지 D'의 신뢰구간(confidence interval)에 따라 가상데이타를 생성하였다.

-D'의 신뢰구간이 [0.8~1.0] 사이인 데이타
-D'의 신뢰구간이 [0.58~0.8] 사이인 데이타
-D'의 신뢰구간이 [0.0 ~0.4] 사이인 데이타

여기에서 D'의 신뢰구간이란 앞에서 언급하였던 D'의 구간추정 범위이다. 신뢰구간의 상한(upper bound)과 하한(lower bound)이 높을수록 강한 연관성을 가진다고 볼 수 있다. 일반적으로 연관불균형의 척도로 사용되는 D'값은 연관성 정도의 점추정(point estimate)값으로 대립유전자의 빈도가 극단적으로 낮거나 혹은 크거나 또는 표본의 수가 작을 때, 연관성이 높은 블록의 구성은 매우 불안정 하다[15]. 따라서 본 실험에서는 데이타의 블록을 D'의 신뢰구간 값으로 나누었고, 신뢰구간을 구하는 방법으로는 붓스트랩(bootstrap) 방법을 사용하였다[16].

2) International HapMap Project 데이타: 22번 염색체(chromosome)에 있는 rs135260~rs130710의 6개의 사이트로 구성되어 있는 사이트간 강한 연관성을 가지는 데이타로 2004월 2월에 HapMap에 발표된 데이타이다[14]. 일반적으로 D'의 신뢰구간 하한이 0.7 이상 상한 0.98이상인 구간이 전체 쌍의 95% 이상되면 사이트간에 강한 연관성을 가진다고 볼 수 있다 [15,17].

본 논문에서 제시한 일배체형 기반의 대치법과 연관불균형 기반의 대치법을 적용하여 기존의 결측치 대치에 많이 쓰이던 주형질 대치법과 비교하여 본 결과가 그림 4, 5, 6 과 같다. 그림 4는 D'의 신뢰구간이 [0.8~1.0]사이인 가상의 데이타에서 각 대치 방법별로 100번의 반복 실험을 한 결과이다. 그림 5와 그림 6은 각각 D'신뢰구간 [0.58~0.8], [0.0~0.4]인 데이타에서 같은 방법으로 실험한 결과이다.

그림 4, 5와 표 4의 결과에서 보듯이 두 개의 데이타 모두에서 주형질 대치법보다 일배체형과 연관불균형 기

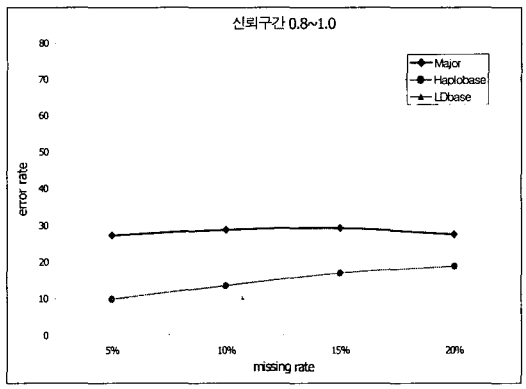

그림 4 D' 신뢰구간 [0.8~1.0]에서의 가상데이타 실험 결과

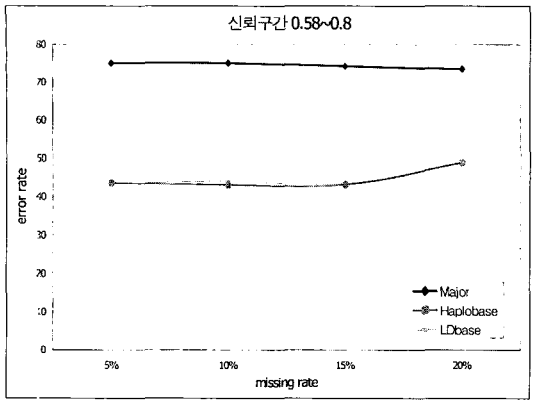

그림 5 D′ 신뢰구간 [0.58~0.8]에서의 가상데이타 실험 결과

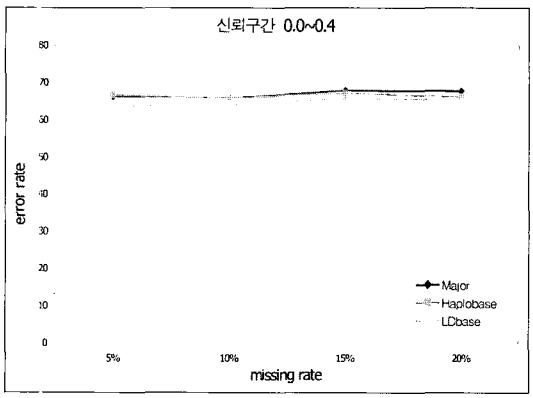

그림 6 D′ 신뢰구간 [0.0~0.4]에서의 가상데이타 실험 결과

반의 대치법이 오차율이 낮았다. D′의 신뢰구간별로 보면 그림 4가 그림 5, 그림 6 보다는 결측치에 대한 오차율이 낮았고 결측치가 증가하면서 오차율도 커지는 안정된 분포를 보임을 알 수 있다. 그러나 신뢰구간이 [0.0~0.4]사이인 데이타의 실험결과인 그림 6을 보면 대치법에 관계없이 높은 오차율을 보임을 알 수 있다. 이는 연관성이 낮은 영역에 대하여는 본 논문에서 제시하는 방법과 기존의 방법간의 성능 차이가 없음을 나타내는데, 일반적으로 이웃하고 있는 단일염기변이 사이트들 간에는 강한 상관관계를 나타내고 있는 것을 감안하면 실제성능에서는 결과에 큰 차이를 보일 것으로 생각된다.

또한 실험의 결과로부터 우리는 연관성이 높은 사이트들의 블록에서 결측치에 대한 오차율이 낮아짐을 볼 수 있었다. 즉, 그림 4와 같은 연관성이 높은 데이타에서의 결측치에 대한 예측률이 더욱 커지고, 기존의 주형질 대치법보다는 본 논문에서 제안하는 두 가지 대치방법이 월등한 예측률을 보이는 것을 볼 수 있었다.

마지막 실험은 실제 International HapMap Project에서 공개된 22번 염색체 상의 강한 연관성을 가지는 6개의 사이트로 구성된 데이타를 가지고 실험하였는데, 그림 7과 표 5의 결과와 같다. 이번 결과도 가상 데이타

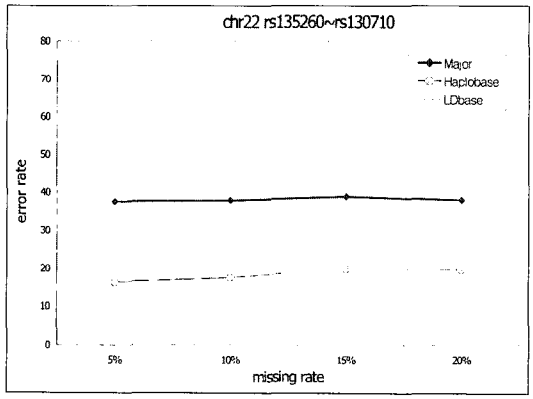

그림 7 International HapMap 데이타에서의 실험결과

표 4 D′ 신뢰구간에 따른 가상 데이타에서의 결측치에 따른 오차율

데이타	결측치율 (%)	주형질대치법	일배체형대치법	연관불균형대치법
신뢰구간 [0.8~1.0]	5	27.31	9.92	9.92
	10	28.85	13.52	13.59
	15	29.30	16.98	17.00
	20	27.57	18.85	18.83
신뢰구간 [0.58~0.8]	5	75.08	43.54	43.31
	10	75.07	43.07	44.11
	15	74.25	43.15	43.40
	20	73.46	48.82	48.80
신뢰구간 [0.0~0.4]	5	66.31	66.77	63.62
	10	65.93	65.82	64.07
	15	67.80	67.15	65.85
	20	67.69	66.22	65.44

표 5 HapMap 데이타에서의 결측치에 따른 오차율

데이타	결측치율 (%)	주형질대치법	일배체형대치법	연관불균형대치법
22번 염색체의 rs135260~rs130710 사이트	5	37.519	16.222	16.593
	10	37.833	17.519	17.815
	15	38.889	19.79	20.247
	20	37.972	19.63	20.065

실험 결과와 마찬가지로 주형질 대치법 보다는 일배체형 기반의 대치법과 연관불균형 기반의 대치법에서 결측치에 대한 오차율이 월등히 낮음을 볼 수 있다.

5. 결론 및 향후 연구 과제

본 논문에서는 생물학 데이타에서 여러 가지 원인으로 인해 결측치가 발생하였을 경우 기존에 일반적으로 쓰이던 주형질 대치법 대신에 새로운 확률값을 기반으로 한 일배체형 기반의 대치법과 연관불균형 기반의 대치법을 제시하였다. 또한 논문에서 제시하는 방법을 사용하여 가상의 데이타와 실제 International HapMap에 공개되어있는 단일염기변이 데이타를 가지고 실험을 하였다. 제시된 연관불균형 방법과 일배체형 방법은 데이타의 사이트간의 연관성이 낮을 때에는 기존에 사용되어 오던 주형질 대치법과 실험 결과가 비슷하였으나 사이트간의 연관성이 높을수록 결측치에 대한 예측률이 높아졌다. 이러한 이유는 앞에서도 언급하였듯이 연관성이 높은 사이트들 사이에는 유전자들의 변이가 제한적으로 나타나는 유전형 자료의 특이성 때문이다.

본 논문에서 제시한 대치방법들은 단일염기변이 같은 유전형 데이타의 자료특이성에 알맞게 만들어져 이러한 자료의 결측치 대치시에 기존의 다른 생물학적 데이타에서 일반적으로 사용되어오던 주형질 대치법보다는 정확한 결측치 대치를 할 수 있다. 하지만 유전형 자료는 사이트들 간의 연관성이 높은 구간이라도 이형접합성(heterozygosity)이나 동형접합성(homozygosity)의 비율 같은 여러 요소에 따라 결측치의 추정 오차율이 변화한다. 향후 이런 다양한 유전형 자료의 특이성을 반영한 좀 더 안정적인 대치 방법론이 필요할 것이다.

참 고 문 헌

[1] John I Bell, "Single nucleotide polymorphisms and disease gene mapping," Arthritis Research, Vol.4, pp.S273-S278, 2002.

[2] Benjamin A. Salisbury, Manish Pungliya, Julie Y. Choi, Ruhong Jiang, Xiao Jenny Sun, and J. Claiborne Stephens, "SNP and haplotype variation in the human genome," Mutation Research, Vol.526, pp.53-61, 2003.

[3] Shin Lin, David J. Cutler, Michael E. Zwick, and Aravinda Chakravarti, "Haplotype Inference in Random Population samples," Am. J. Hum. Genet.., Vol.71, pp.1129-1137, 2002.

[4] Lon R. Cardon and Goncalo R. Abecasis, "Using haplotype blocks to map human complex trait loci,"Trends in Genetics, Vol.19, pp.135-140, 2003.

[5] Young-sool Park and Soon-kwi Kim, "Comparative Study on Imputation Procedures in Exponential Regression Model with missing values," Journal of Korean Data & Information Science Society, Vol.14, pp.143-152, 2003.

[6] Hyun-Jeong Kim, Sung-Ho Moon, and Jae-Kyoung Shin, "Application of NORM to the Multiple Imputation for Multivariate Missing Data," Journal of Korean Data & Information Science Society, Vol.13, pp.105-113, 2002.

[7] Sung-Ho Moon, Hyun-Jeong Kim, and Jae-Kyoung Shin, "Application of SOLAS to the Multiple Imputation for Missing Data," Journal of Korean Data & Information Science Society, Vol.14, pp.579-590, 2003.

[8] M. Carol, et al., "A comparison of Imputation Techniques for Handling Missing Data," Western Journal of Nursing Research, Vol.24, pp.815-829, 2002.

[9] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman, "Missing value estimation methods for DNA microarrays," Bioinformatics, Vol.17, pp.520-525, 2001.

[10] Anthony J. Brookes, "The essence of SNPs," Gene, Vol.234, pp.177-186, 1999.

[11] Zhaohui S. Qin, Tianhua Niu, and Jun S. Liu, "Partition-Ligation-Expectation-Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms," Am. J. Hum. Genet. Vol.71, pp.1242-1247, 2002.

[12] R. C. Lewontin, "The interaction of selection and linkage. I. General considerations; heterotic models," Genetics, Vol.49, pp.49-67, 1964.

[13] http://www.people.fas.harvard.edu/~junliu/plem/click.html/.

[14] http://www.hapmap.org/index.html.en/.

[15] Stacey B. Gabriel, et al., "The structure of haplotype blocks in the human genome," Science,

Vol.296, pp.2225-2229, 2002.
[16] B. Efron, "Bootstrap methods: another look at the jackknife," Ann. Stat., Vol.7 pp.1-26, 1979.
[17] Thomas G. Schulze, Kui Zhang, Yu-Sheng Chen, Nirmala Akula, Fengzhu Sun, and Francis J. McMahon, "Defining haplotype blocks and tag single nucleotide polymorphisms in the human genome," Human Molecular Genetics, Vol.13, pp.335-342, 2004.
[18] B. Devin and Neil Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping", Genomics, Vol.29, pp. 311-322, 1995.

박 윤 주

1999년 동덕여자대학교 전산통계학과 졸업(학사). 2001년 연세대학교 의학통계학과 졸업(이학석사). 2001년~2002년 포항공대 생물학 전문 연구정보센터(BRIC) 2003년~현재 국립보건연구원 유전체연구부 기술전문연구원. 관심분야는 생물정보학, SNP 분석

김 영 진

2003년 한동대학교 생명식품과학부 졸업(학사). 2003년 7월~현재 국립보건연구원 유전체연구부 연구원. 관심분야는 생물정보학, DNA chip 분석

박 정 선

2001년 한동대학교 생물식품공학부 졸업(학사). 2003년 한동대학교 정보통신학과 졸업(공학석사). 2002년 10월~현재 국립보건연구원 유전체연구부 기술전문연구원. 관심분야는 생물정보학, SNP 분석

김 규 찬

의학박사/전문의. 미국 소아과 전문의/미국 면역학 전문의. 1975년 서울대학교 의과대학 졸업(의학사). 1979년 서울대학교 의과대학 졸업(의학석사). 1982년 서울대학교 의과대학 졸업(의학박사). 1980년~1981년 한림대의대 강남성심병원(부과장). 1981년~1984년 아산재단 금강병원(과장). 1984년~1986년 Cornell대 의대 및 Sloan Kettering 암센터 1986년~1989년 UCLA 의과대학 소아과/면역분과 (Post-doc Fellow). 1989년~1990년 UCLA 의과대학 면역연구센터 및 임상면역연구실(Research Scholar). 1990년~1992년 UCLA 의과대학 UCLA-Harbor 의료원(조교수). 1992년~1996년 인제대학교 백병원(교수, 과장). 1996년~2000년 국립보건원 특수질환부(과장). 2001년~ 현재 질병관리본부 국립보건연구원 유전체연구부(운영위원장/부장). 관심분야는 임상의학, 유전체학

고 인 송

1987년 고려대학교 의과대학 졸업(의학사). 1993년 Boston University, Dept. of Cognitive and Neural Systems졸업(석사: 뇌공학전공: M.A. in Neural Networks). 1993년~1997년 Boston University, Dept. of Cognitive & Neural Systems박사 수료 (뇌과학 전공: Computational Neuroscience). 2001년 고려대학교 의과대학 졸업(의학박사: 생리학전공). 1988년~1989년 중앙대학교 의과대학 생리학 조교(Electrophysiology 전공). 1989년~1990년 George Washington Univ, Computer Science Dept. 연수. 1998년1월~2002년 2월 국립보건원 생명의학부/유전체연구소 보건연구관. 2002년 3월~2003년 8월 고려대학교 의과대학 생명정보학 연구교수. 2003년 9월~2004년 11월 국립보건연구원 유전체연구부 생명정보학연구단장. 2004년 11월~현재 과학기술혁신본부 기술혁신평가국 보건연구관. 관심분야는 생물정보학, 의료정보학, 수리생물학, 신경정보학

정 호 열

1997년 부산대학교 전자계산학과 졸업(학사). 1999년 부산대학교 전자계산학과 졸업(이학석사). 2002년 부산대학교 전자계산학과 졸업(이학박사). 2002년 11월~2004년 5월 국립보건연구원 유전체연구부 책임전문연구원. 2004년 6월~현재 한국전자통신연구원 바이오정보연구팀 선임연구원. 관심분야는 생물정보학, 조절 네트워크, 유전자 클러스터링, 그래프 이론, 이미지 처리