iHaplor: A Hybrid Method for Haplotype Reconstruction

Ho-Youl Jung[1], Jee-Yeon Heo[2], Hye-Young Cho[1], Gil-Mi Ryu[1], Ju-Young Lee[1], InSong Koh[1], Kuchan Kimm[1], Bermseok Oh[1*]

[1] National Genome Research Institute, NIH, Seoul, Korea
[2] The Sensory Research Center, College of Pharmacy, Seoul National University

*To whom correspondence should be addressed. E-mail: ohbs@nih.go.kr

Abstract

This paper presents a novel method that can identify the individual's haplotype from the given genotypes. Because of the limitation of the conventional single-locus analysis, haplotypes have gained increasing attention in the mapping of complex-disease genes. Conventionally there are two approaches which resolve the individual's haplotypes. One is the molecular haplotypings which have many potential limitations in cost and convenience. The other is the *in-silico* haplotypings which phase the haplotypes from the diploid genotyped populations, and are cost effective and high-throughput method. *In-silico* haplotyping is divided into two sub-categories - statistical and computational method. The former computes the frequencies of the common haplotypes, and then resolves the individual's haplotypes. The latter directly resolves the individual's haplotypes using the perfect phylogeny model first proposed by Dan Gusfield [7]. Our method combines two approaches in order to increase the accuracy and the running time. The individuals' haplotypes are resolved by considering the MLE (Maximum Likelihood Estimation) in the process of computing the frequencies of the common haplotypes.

Introduction

SNP(Single Nucleotide Polymorphism)은 집단유전학(population genetics) 연구와 complex disease 에 영향을 주는 변이에 대한 마커(marker)로 유용하게 사용될 수 있다. 이런 목적으로 사용되어 온 다른 마커들이 있으나, 현재까지 수백만 개의 SNP 이 발견되어서 인간 유전체에 dense 한 분포를 보이고 있으므로 SNP 을 이용하면 질병과의 연관성(association) 연구에 있어서 dense marker 로 사용될 수 있다는 장점으로 많이 사용되고 있다. 질병군과 정상군간의 유전학적인 변이를 살펴보는 것으로 SNP 과 질병과의 연관성(association)을 조사하는 방법을 사용하는 경우가 있으나, 각 개인의 모든 SNP site 에 대한 genotyping 비용이 많이 들어, LD(Linkage Disequilibrium) block 별로 haplotype 을 구하여 질병 연관성을 분석하는 방법을 많이 사용하고 있다 [4]. Haplotype 을 이용하여 질병 연관성을 분석하는 경우에는 인구 집단 내에서 각 개인별 haplotype 을 찾아 주는 방법이 필요로 하고, 이런 과정을 일반적으로 haplotype reconstruction 문제라고 한다. Halotype reconstruction 은 크게

- 생물학적 접근 방법(molecular method);
- 계통적 접근 방법(direct inference from family data);
- 계산적 접근 방법(*in-silico* method);

의 3 가지로 나누어 생각할 수 있다 [10].

This study was supported by the intramural grant of the National Institute of Health, Korea.

생물학적인 접근 방법의 대표적인 것으로 allele-specific PCR 을 예로 들 수 있는데, 어느 정도 high-throughput 은 보장하나 비교적 짧은 haplotype 에 적당한 방법이고 실험에 들어가는 비용도 비교적 많은 편이다. 그리고 Patil 등이 21 번 염색체에서의 haplotype 변이를 분석하기 위해 사용한 rodent-human somatic cell hybrid 에서 해당 염색체를 haploid 로 나누어 직접 haplotype 을 찾아주는 방법이 있으나, 현재까지 그 실험의 용이성이 높지 않으며 그 비용 또한 많이 들어가는 단점을 가지고 있다 [10].

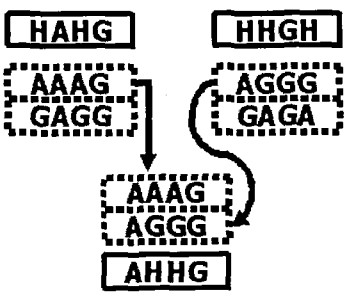


그림 1: 계통적 접근 방법에 의한 haplotype reconstruction 의 한 예: 아버지는 HAHG(AAAG, GAGG), 어머니는 HHGH(AGGG, GAGA), 그리고 자식은 AHHG(AAAG, AGGG)의 genotype(haplotype pair)을 가지고 있다.

계통적 접근 방법은 family data 로부터 haplotype 을 찾아주는 것으로 기본적으로 부, 모, 자식의 genotype data 를 하나의 그룹으로 하여 각 개인의 haplotype 을 계산하는 방법이다. 그림 1 은 하나의 예로써, 네 개의 이어진 SNP 에 대해서 가능한 염기가 A 와 G 라 하고 heterozygous site 는 H 라고 할 때, 자식의 genotype 이 AHHG, 어머니는 HHGH 그리고 아버지는 HAHG 라고 하자. 이 경우에 3 명의 genotype 을 입력으로 받아서 각각의 haplotype 은 계산을 통해서 알 수가 있는데, 자식의 첫 번째 SNP site 의 genotype 이 A 이므로 어머니와 아버지로부터 전달된 두 개의 haplotype 의 첫 번째 SNP site 모두 A 임을 알 수 있다. 따라서 어머니와 아버지로부터 전달되지 않은 haplotype 의 첫 번째 SNP site 는 G 임을 쉽게 알 수 있다. 그리고 자식의 두 번째 SNP site 는 H 로 heterozygous 이고 아버지는 homozygous 이므로 A 는 아버지로부터 G 는 어머니로부터 전달 받았음을 알 수 있다. 이 과정을 4 개의 SNP site 에 반복적으로 적용하면 자식의 haplotype 은 (AAAG, AGGG), 어머니의 haplotype 중 자식에게 전달된 것은 AGGG, 전달되지 않은 것은 GAGA, 그리고 아버지의 haplotype 중 자식에게 전달된 것은 AAAG, 전달되지 않은 것은 GAGG 이다. 이 방법은 입력되는 genotype data 가 극단적인 경우가 아닌 경우에는 신뢰성이 있는 결과를 얻을 수 있으나, data 를 수집하는 것이 쉽지 않으며 계산적 접근 방법을 사용하는 경우보다 그 비용이 많이 소요된다고 할 수 있다 [5].

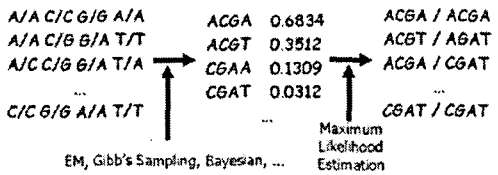


그림 2: Genotype data 를 입력으로 받아 각 개인별 haplotype 을 생성하는 계산적 접근 방법의 일반적인 형태.

계산적 접근 방법은 인구 집단의 genotype data 를 입력으로 하여 각 개인의 haplotype 을 생성하는 것으로, 기본적으로 통계적인 개념에 근거를 두고 있다. SNP site 가 M 개인 영역에 대해서 인구 집단의 haplotype 분포를 조사하였을 때, 인구 집단 내에 많이 나타나는 공통의 haplotype (common haplotype)이 일반적으로 M+1 개 이하로 나타나는 특징에 기초하고 있다 [4]. 일반적인 과정은 그림 2 와 같이 genotype data 에서 common haplotype 이 발생할 수 있는 빈도수(frequency)를 추정하여 이것을 기초로 각 개인의 haplotype 을 MLE(Maximum Likelihood Estimation)를 이용하여 계산하는 방법을 사용하고 있다 [2, 3, 6, 8, 9, 11, 12]. 공통 haplotype 의 빈도수를 추정하는 데 있어 EM(Expectation Maximization), Gibbs sampling, 또는 Bayesian 등과 같은 방법을 사용하고 있다. 공통 haplotype 의 빈도수를 추정하는 여러 방법들은 어느 정도 신뢰성이 있는 결과를 얻을 수 있으나, 수행 시간이 오래 걸리는 단점과 어떤 생물학적인 의미도 가지고 있지 않다는 단점을 가지고 있다. 한편 이런 통계적인 접근 방법이 아니라 나타날 수 있는 haplotype 을 tree 에 mapping 하는

perfect phylogeny 방법을 이용한 것도 있다. 이는 haplotype 이 어느 정도 진화의 형태를 가지고 있다는 가정하에 계통적인 개념에서 접근한 것으로 생물학적인 의미를 가지고 있으나, perfect phylogeny 를 만족하지 못하는 경우에는 해결 방법을 제시하지 못하는 단점을 가지고 있다 [7]. 이런 문제를 해결하기 위해 Eskin 등은 수정된 모델을 제시하였으나, heuristic 을 사용하여 본질적인 해결 방법은 아니다 [5].

그리고, 계산적 접근 방법에서 통계적인 개념을 바탕으로 하는 경우에는 그림 2에서 보듯이 genotype data에서 공통의 haplotype을 추정하는 단계와 MLE(Maximum Likelihood Estimation)하는 단계에서 오차가 발생할 가능성이 있다. MLE는 피할 수 없는 단계라 하면 이 오차를 줄이기 위해서는 공통 haplotype의 빈도수를 추정하는 단계에서 오차를 줄여야 각 개인별로 더 정확한 haplotype 추정이 가능할 것이다. 따라서 본 논문에서는 공통 haplotype의 빈도수를 추정하는 단계에서 기존의 통계적인 추정치를 사용하는 것이 아니라, haplotype의 진화성에 기초하여 더 정확한 추정치를 계산할 수 있는 방법과 이 추정치를 MLE에 적용하여 각 개인별 haplotype을 선형 시간에 계산하는 새롭고 수행시간이 빠른 방법론을 제시한다.

Methods

Preliminaries

논문에서는 수학적 표현 방법의 통일성을 위하여 Gusfield [7]가 사용한 표현 방법을 같이 사용한다. Haplotype reconstruction을 수행하기 위해 입력으로 들어오는 데이터는 genotype matrix G 로 n개의 genotype vector로 구성되어 있으며, 각 vector의 길이를 m이라 하면 genotype은 G_{ij}, $1 \le i \le n$, $1 \le j \le m$ 로 나타낸다. 그리고, haplotype reconstruction의 결과는 genotype matrix G 에 해당하는 haplotype matrix H 로 n개의 haplotype vector pair로 genotype G_{ij} 에 대한 haplotype은 각각 h_{ij}^1, h_{ij}^2로 표현할 수 있다. 즉 입력으로 들어오는 개인의 genotype data의 개수는 n이고 한명의 genotype은 m개의 SNP site에 대한 genotyping 결과를 의미한다. G_{ij} 의 값은 0, 1, 또는 2의 값을 가지고, 0, 1은 homozygous site를 의미하며, 2는 heterozygous site를 각각 의미한다. Haplotype reconstruction 문제는 입력으로 주어진 n개의 genotype vector를 $2 \cdot n$ 개의 이진 vector로 decompose하는 것과 같은 문제이다. 이런 과정을 일반적으로 phasing 혹은 resolving이라고 하며 $g_{ij} = 0,1$ 이면 해당하는 haplotype pair는 $h_{ij}^1 = h_{ij}^2$ 이고, $g_{ij} = 2$ 이면 $h_{ij}^1 \neq h_{ij}^2$ 이다. 여기서 만약에 어떤 genotype vector g_i 에서 heterozygous site의 개수를 n_i^h 라 하면 가능한 haplotype의 개수는 $2^{n_i^h}$ 이다. 이렇듯 haplotype reconstruction에서 하나의 genotype이 여러 개의 해를 가질 수 있는 문제를 phasing problem, 혹은 resolving problem이라고 한다. 예를 들어, 어떤 genotype vector $g_i =< 0,2,1,2 >$ 라고 하면, haplotype vector pair, <0,1,1,0>과 <0,0,1,1>이 하나의 해가 될 수 있으며, <0,0,1,0>과 <0,1,1,1>도 다른 하나의 해가 될 수 있는 문제를 가지고 있다. 이러한 phasing problem을 해결하기 위해서 하나의 genotype vector만 고려해서 phasing 작업을 수행하는 것이 아니라, 전체 genotype vector를 모두 고려하여 확률적으로 높은 haplotype pair를 각각의 genotype vector에 적용하는 방법을 사용한다.

Hybrid method

이전에서 소개한 바와 같이 haplotype reconstruction 방법 중에서 통계적 접근 방법은 그림 2와 같이 입력된 전체 genotype에서 빈번히 나타나는 haplotype-일반적으로 common haplotype이라고 함-들의 빈도를 추정한다. 이 common haplotype frequency를 기반으로 MLE(Maximum Likelihood Estimation)를 적용하여 각 genotype에 대한 haplotype pair를 계산하는 방법을 사용하고 있다. Common haplotype frequency는 가능한 모든 haplotype들에 대한 frequency를 모두 계산하는 경우에는 $\Theta(2^m)$ 개의 memory 공간이 필요로 하기 때문에 실제적으로 모든 패턴을 계산하는 것은 불가능하고 이와 같이 추정에 의한 방법을 사용할 수 밖에 없다. 그리고 haplotype의 계통적인 측면을 고려한 perfect phylogeny에 기반한 방법들은 중간의 common haplotype frequency를 추정하지 않고,

각 genotype에서 가능한 haplotype들 중에서 perfect phylogeny를 만족할 수 있는 haplotype pair로 phase해주는 방법을 사용하고 있기에 수행 속도가 빠른 장점이 있는 반면에 입력된 genotype들에서 가능한 haplotype들이 perfect phylogeny를 만족하지 않는 경우에는 haplotype reconstruction을 수행 할 수 없는 단점을 가지고 있다. 따라서 본 논문에서는 두 가지 방법의 단점을 보완할 수 있도록, haplotype의 계통적인 측면을 고려하여 common haplotype frequency를 계산하고 계산된 결과를 MLE를 적용하여 각 genotype에 대한 haplotype pair를 선형 시간에 계산하는 방법을 제시한다. 따라서 본 논문에서 가정하고 있는 조건은 다음과 같다.

- 가정 1: recurrent mutation은 없다.
- 가정 2: 입력되는 genotype은 haplotype block 범위, 즉 recombination이 없는 영역에 대한 것 이다.
- 가정 3: mutation은 SNP site에서 최소 변이를 보이는 것으로부터 유도된 것이다.

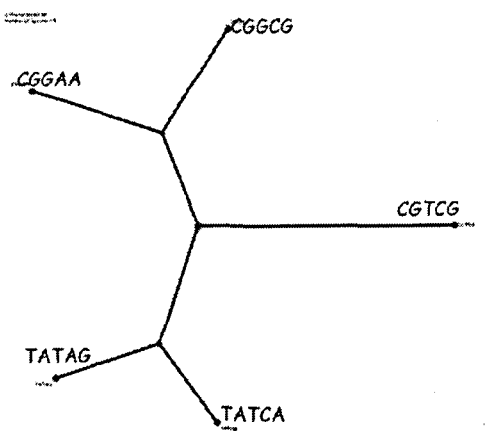


그림 3: Haplotype의 계통도: Daly의 데이터 중에서 3번째 block내의 common haplotype들에 대한 계통도이다. 계통도는 PhyloDraw [1]로 가시화하였다.

가정 1과 2는 haplotype reconstruction을 수행하는 다른 방법들에서 가정하고 있는 것이고 이러한 가정이 없다면 *in-silico*방식으로 haplotype을 계산하는 것이 의미가 없거나 부정확한 결과일 수 있다. 가정 3에 대한 것은 그림 3에서 보듯이 common haplotype들을 계통적인 측면에서 보면 서열이 비슷하면 계통도에서 거리가 가깝고, 서열의 차이가 많으면 거리도 먼 것을 쉽게 알 수 있다. 예를 들어, CGGAA는 CGGCG와는 2개의 SNP site에서 차이를 보이고 있고, TATAG와는 4개의 SNP site 차이를 보이고 있어 계통도상에서 TATAG보다 CGGCG와 더 가깝게 나타남을 알 수 있다. 이를 통해서 CGGAA는 TATAG보다는 CGGCG로부터 유도되었을 가능성이 더 높다는 것을 알 수 있다.

본 논문에서 제시하는 haplotype reconstruction 알고리즘은 먼저 genotype matrix G를 입력으로 받아서 g_i를 n_i^h 값에 따라 정렬하여 이 값을 기준으로 G'과 G''의 2개의 group으로 양분한다. 그러면, G'에 속한 genotype vector들은 모두 n_i^h가 1이하로 phasing problem없이 haplotype pair를 쉽게 알 수 있고 그 결과를 haplotype matrix H'으로 생성한다. 예를 들어, <0,1,0,2>와 같은 genotype vector는 <0,1,0,0>과 <0,1,0,1>의 haplotype vector pair로 쉽게 phasing할 수 있음을 알 수 있다. 그리고 G''에 속하는 genotype vector는 가정 3을 기준으로 H'에 속하는 haplotype vector들 중에서 가장 가깝고 phasing 규칙에 부합되는 haplotype vector pair로 phasing하여 H''을 생성한다.

예를 들어, 입력으로 주어진 genotype matrix, G가 그림 4처럼 {TATCGT, TACGC, CCACH, HACGC, HACGT, HACGH}로 구성되어 있다고 하자. 그러면 {TATCGT, TACGC, CCACH, HACGC, HACGT}는 heterozygous site(H)의 개수가 1이하이므로 이들 genotype은 G'의 원소가 되고 이들을 phasing한 haplotype은 H'의 원소들 ({TACGT, TACGC, CCACT, CCACC, CACGC, CACGT})이 된다. 그리고, {HACGH}는 heterozygous site(H)의 개수가 2이상이므로 G''의 원소가 되고 {HACGH}에서 heterozygous site를 제외한 나머지 site들과 H'에 있는 haplotype들을 비교하여 같은 haplotype을 찾으면, {(TACGT, CACGC), (TACGC, CACGT)}가 가능하고 괄호로 묶여 있는 것은 HACGH에서 phasing rule을 만족하는 pair를 의미한다. 두 개의 pair에 대해서 이런 haplotype pair가 나타날 수 있는 확률(P_h)을 allele frequency를 이용하여 다음과 같이 계산한다.

$$P_h = \min\{\Pi p(h_i^j), 1 \le i \le m \text{ and } j = 1,2\},$$

여기서, $p(h_i^j)$ 은 i번째 SNP site의 h_i^j allele frequency를 의미하고, m은 SNP site의 개수를 의미하고 $j = 1,2$ 는 haplotype pair 중에서 첫 번째와 두 번째 haplotype을 각각 의미한다. 두 개의 haplotype pair에 대해서 각각 계산한 확률, P_h 가 높은 pair(본 논문에서는 가상으로 (TACGT, CACGC) pair라고 함)를 선택하여 H'' 의 원소로 삽입한다. 여기서 각 haplotype pair에 대해서 나타날 수 있는 확률 값이 최소인 것이 최대가 되는 것을 선택하는 이유는 MLE에서와 유사하게 균일하게 나타나는 haplotype pair를 선택하는 것이 더 좋은 결과를 주기 때문이다. G'' 의 모든 원소에 대해서 위 과정이 완료되면 H' 과 H'' 을 합해서 haplotype matrix, H 를 생성한다.

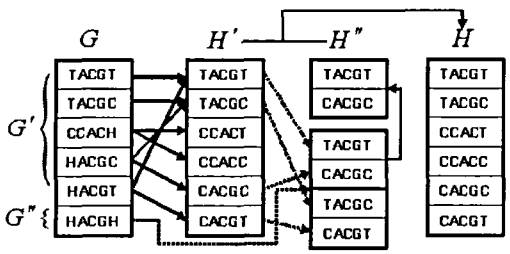


그림 4: Genotype matrix, G 로부터 haplotype matrix, H 를 계산하는 예.

Haplotype reconstruction의 전체 알고리즘은 알고리즘 1과 같다. 알고리즘 2는 ``phasing_ambiguous( $G''$ )"에 대한 구체적인 알고리즘이다. 여기서, heterozygote( $g_i$ )가 의미하는 것은 genotype $g_i$ 중에서 heterozygous site, 즉 $g_{ij} = 2$ 인 SNP site들을 의미하는 것이다. 그리고, ``search exact matching haplotypes (H', heterozygote(g_i))"은 genotype vector, g_i 에서 heterozygous site를 제외한 것과 H' 에 있는 haplotype vector 중에서 일치하는 모든 것을 찾는 함수를 의미한다.

Algorithm: Haplotype reconstruction
Input: genotype matrix, G
Output: haplotype matrix, H

```
for i = 1 to n
  if n_i^h ≤ 1 then G' ← g_i ;
  else G'' ← g_i ;
endfor
H' ← phasing_unambiguous(G') ;
H'' ← phasing_ambiguous(G'') ;
combine H' and H'' ;
```

알고리즘 1: 여기서 n_i^h 는 genotype vector내에 heterozygous site의 개수를 의미한다. " phasing_unambiguous(G') "는 heterozygous site가 없거나 혹은 1개이기에 phasing 규칙에 의해서 쉽게 haplotype pair를 구할 수 있다.

Results

본 논문에서 사용한 실험 데이터는 Daly 등에 의해서 생성된 129개의 계통 데이터이다 [3]. 이 데이터에서 인구 집단의 총 수는 n = 129이며 SNP site의 개수는 m = 103이다. 그리고 103개의 SNP site는 총 11개의 LD(Linkage Disequilibrium) block내에서 각각 수행하였다. 표 1은 2번째 block에 대해서 수행한 결과를 PL-EM [11]과 비교하여 보여주고 있다. 표 1의 첫번째 열은 입력으로 주어지는 genotype들을 나타내고 두번째 열은 결과로 나와야 되는 true haplotype pair를 나타내고 있다.

Haplotype reconstruction과 관련하여 이전에 제시되었던 많은 방법들과 비교-분석을 수행하기 위해서 본 논문에서는 두가지 에러 측정 방법을 제시한다. 첫째는 individual phasing error(e_p)로 이는 인구 집단내의 어떤 genotype에 대해서 phasing을 잘못한 경우 1증가하는 에러 측정 방법이다. 둘째는 base phasing error(e_b)로 haplotype matrix, H 전체에서 SNP base 단위로 true haplotype과 비교하여 잘못되어진 phasing이 있으면 e_b 를 1증가하는 방법이다. 만약에 두개의 haplotype reconstruction하는 방법론, HR_A와 HR_B가 있다고 할 때, HR_A와 HR_B가 똑 같은 e_p (혹은 e_b) 값을 가진다면 둘

중에서 e_b (혹은 e_p)가 낮은 것이 좋은 방법이라고 할 수 있을 것이다. 그래서 본 논문에서는 기존에 많이 사용하고 있는 Haplotyper [12]와 PL-EM[11]과의 성능 비교를 하였고 표 2에 그 결과가 나와있고, 그림 5에서 11개의 block 데이터에 대해서 e_p 와 e_b 의 값을 시각적으로 표현하여 비교하였다. 그 결과로 본 논문에서 제시하는 방법이 더 좋은 결과를 보임을 쉽게 알 수 있다.

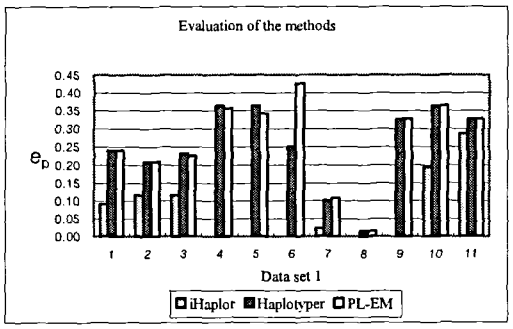
Data	iHaplor		Haplotyper		PL-EM	
	e_p	e_b	e_p	e_b	e_p	e_b
Block1	0.09	0.02	0.24	0.06	0.24	0.06
Block2	0.12	0.03	0.22	0.05	0.20	0.05
Block3	0.12	0.06	0.23	0.16	0.22	0.16
Block4	0.00	0.00	0.36	0.10	0.36	0.10
Block5	0.00	0.00	0.36	0.07	0.34	0.07
Block6	0.00	0.00	0.25	0.07	0.43	0.09
Block7	0.02	0.01	0.10	0.03	0.11	0.03
Block8	0.00	0.00	0.02	0.01	0.02	0.01
Block9	0.00	0.00	0.33	0.11	0.33	0.11
Block10	0.19	0.05	0.36	0.09	0.36	0.09
Block11	0.29	0.14	0.33	0.15	0.33	0.15
Average	0.08	0.03	0.25	0.07	0.27	0.07

표 2: 본 논문에서 제시하는 방법과 다른 방법(Haplotyper [12], PL-EM [11])과의 비교-분석 결과.

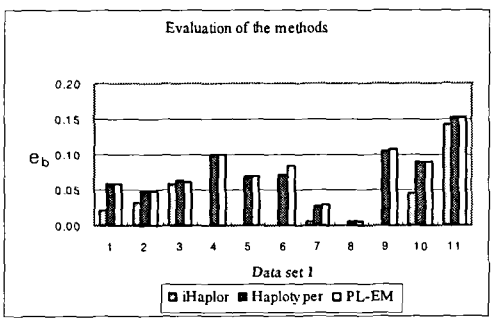
Discussion

본 논문에서는 genotype과 phenotype과의 연관성을 보는 association 연구의 기초가 되는 SNP site에 대한 인구 집단의 genotype data에서 각 개인별 haplotype을 *in-silico* 방식으로 계산하여 주는 방법론을 제시하였다. 제시한 방법론은 이전의 연구와 비교하여 그 수행 속도와 정확성에 있어서 더 좋은 결과를 보여주고 있다. 수행 속도는 인구 집단의 크기가 N이라고 하면 $O(N)$의 선형 시간에 계산이 가능하며, 정확성은 실제 데이터와 가상의 데이터 모두에 대해 평균적으로 92% 정도를 보이고 있다. 향후 연구과제로서,

- 실제 데이터에서는 genotype에서 missing이 되는 값이 있는데 이에 대한 imputation module이 필요하다;
- Haplotype은 LD block영역 내에서 계산하는 것이 의미를 가지므로 주어진 genotype 데이터로부터 LD (혹은 haplotype) block을 찾아주는 방법의 구현이 필요하다;
- 계산되어진 haplotype에 대한 가시화 방법이 필요하다.



(a) 세 가지의 방법으로 e_p 를 11개 block에 대해서 계산한 결과



(b) 세 가지의 방법으로 e_b 를 11개 block에 대해서 계산한 결과

그림 5: 본 논문에서 제시하는 방법과 다른 방법과의 비교를 시각적으로 표현.

References

[1] Jeong-Hyeon Choi, Ho-Youl Jung, Hye-Sun Kim, and Hwan-Gue Cho. PhyloDraw: a phylogenetic tree drawing system, *Bioinformatics*, 16(11):1056-1058, 2000.

[2] Andrew G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations, *Molecular Biology and Evolution*, 7(2):111-122, 1990.

[3] Andrew G. Clark, Kenneth M. Weiss, Deborah A. Nickerson, Scott L. Taylor, Anne Buchanan, Jari Stengard, Veikko Salomaa, Erkki Vartiainen, Markus Perola, Eric Boerwinkle, and Charles F. Sing. Haplotype structure and population genetic inferences

from nucleotide-sequence variation in human lipoprotein lipase, *American Journal of Human Genetics*, 63(2):595-612, 1998.

[4] Mark J. Daly, John D. Rioux, Stephen F. Schaffner, Thomas J. Hudson, and Eric S. Lander. High-resolution haplotype structure in the human genome, *Nature Genetics*, 29(2):151-158, 2001.

[5] Eleazar Eskin, Eran Halperin, and Richard M. Karp. Large scale reconstruction of haplotypes from genotype data, In *Proceedings of the Seventh annual International Conference on Computational Molecular Biology (RECOMB 2003)*, pages 104-113, 2003.

[6] Daniele Fallin and Nicholas J. Schork. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data, *American Journal of Human Genetics*, 67(4):947-959, 2000.

[7] Dan Gusfiled. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions, In *Proceedings of the Sixth International Conference on Computational Molecular Biology (RECOMB 2002)*, pages 166-175, 2002.

[8] M.E. Hawley and Kenneth K. Kidd. Haplo: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes, *Journal of Heredity*, 86(5):409-411, 1995.

[9] Tianhua Niu, Zhaohui S. Qin, Xiping Xu, and Jun S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, *American Journal of Human Genetics*, 70(1):157-169, 2002.

[10] Nila Patil, Anthony J. Berno, David A. Hinds, Wade A. Barrett, Jigna M. Doshi, Coleen R. Hacker, Curtis R. Kautzer, Danny H. Lee, Claire Marjoribanks, David P. McDonough, Bich T. N. Nguyen, Michael C. Norris, John B. Sheehan, Naiping Shen, David Stern, Renee P. Stokowski, Daryl J. Thomas, Mark O. Trulson, Kanan R. Vyas, Kelly A. Frazer, Stephen P. A. Fodor, and David R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21, *Science*, 294(23):1719-1723, 2001.

[11] Zhaohui S. Qin, Tianhua Niu, and Jun S. Liu. Partition Ligation-Expectation Maximization Algorithm for Haplotype Inference with Single-Nucleotide Polymorphisms, *American Journal of Human Genetics*, 71(5):1242-1247, 2002.

[12] Matthew Stephens, Nicholas J. Smith, and Peter Donnelly. A new statistical method for haplotype reconstruction from population data, *American Journal of Human Genetics*, 68(4):978-989, 2001.

Algorithm: phasing_ambiguous
Input: genotype matrix, G''
Output: haplotype matrix, H''
for $i = 1$ **to** $|G''|$

$H^0 \leftarrow$ search exact matching haplotypes (H' , heterozygote(g_i));

case $|H^0| = 0$:

$\{h^0\} \leftarrow$ phasing all possible haplotypes from g_i ;

compute the haplotype probability of each haplotype pair in $\{h^0\}$;

add a maximum haplotype probability pair into H'' ;

case $|H^0| = 1$:

$h^0 \leftarrow$ compute a haplotype according to the phasing rule;

add $\{h^0\}$ into H'' ;

case $|H^0| \geq 2$:

$\{h^0\} \leftarrow$ search a haplotype pair coinciding with phasing rule;

if $|\{h^0\}| = 2$ **then**

add $\{h^0\}$ into H'' ;

else

compute the haplotype probability, P_h of each haplotype pair in $\{h^0\}$;

add a maximum haplotype probability pair into H'' ;

endfor

알고리즘 2: heterozygote(g_i)가 의미하는 것은 genotype g_i 중에서 heterozygous site, 즉 $g_{ij} = 2$ 인 SNP site들을 의미한다.

Genotype	True Haplotypes	개수	Our method		PL-EM [11]	
			Haplotype 1	Haplotype 2	Haplotype 1	Haplotype 2
CCAAC	CCAAC/CCAAC	1	CCAAC	CCAAC	CCAAC	CCAAC
TACGT	TACGT/TACGT	18	TACGT	TACGT	TACGT	TACGT
TACGC	TACGC/TACGC	36	TACGC	TACGC	TACGC	TACGC
TACGH	TACGT/TACGC	38	TACGT	TACGC	TACGT	TACGC
CCAAH	CCAAT/CCAAC	1	CCAAT	CCAAC	CCAAT	CCAAC
HACGC	CACGC/TACGC	1	CACGC	TACGC	CACGC	TACGC
HACGT	CACGT/TACGT	1	CACGT	TACGT	CACGT	TACGT
HACGH	**CACGC/TACGT**	**1**	**CACGC**	**TACGT**	**CACGT**	**TACGC**
TACHH	**TACGT/TACAC**	**1**	**TACGT**	**TACAC**	**TACGC**	**TACAT**
HHCHC	CCCAC/TACGC	2	CCCAC	TACGC	CCCAC	TACGC
HHHHC	CCAAC/TACGC	20	CCAAC	TACGC	CCAAC	TACGC
HHHHH	CCAAC/TACGT	9	CCAAC	TACGT	CCAAC	TACGT

표 1: Daly 데이터 [4] 중에서 두 번째 블록에 대한 수행 결과이다. 여기서 genotype HACGH와 TACHH에 대해서 본 논문에서 제시하는 방법이 기존의 다른 방법과 비교하여 더 좋은 결과를 보임을 알 수 있다.