유전자 알고리즘을 이용한 하플로타입 추론

이시영*, 김희철
한국외국어대학교 컴퓨터공학과
e-mail: synuri@hitel.net* , hckim@hufs.ac.kr

Haplotype Inference Using Genetic Algorithm

See Young Lee*, Hee-Chul Kim
Dept. Computer Science of Engineering, Hankook University of Foreign Studies

요 약

사람들 사이에는 DNA 서열의 변이로 인한 유전적 차이가 있으며, 가끔 이러한 차이가 유전 질병의 원인이 되기도 한다. 일반적으로 DNA에서 가장 잘 알려진 변이가 바로 SNP(Single Nucleotide Polymorphism : 스닙)이다. SNP는 보통 블록단위로 유전되어지며 한쪽 부모로부터 유전되어진 SNP 블록을 SNP 하플로타입이라고 부른다. 생물학 실험을 통하여 추출되어진 결과물은 부모로부터 유전되어진 대립 유전자가 혼합되어진 지노타입(genotype)의 정보이다. 지노타입은 직관적으로 정확한 SNP 하플로타입을 추정하기가 힘들고, 생물학 실험을 통하여 하플로타입(haplotype)을 분석하는데 많은 비용이 들기때문에, 이를 컴퓨터 계산을 통하여 추론하는 연구가 Clark[1]에 의해서 제안되어진 이후 활발하게 진행되고 있다. 본 논문에서는 하플로타입을 효과적으로 추론하기 위해 유전자 알고리즘을 이용한 새로운 방법을 설명하고, 실험 결과를 기존의 연구 결과와 비교 분석한다.

1. 서론

SNP는 인류가 진화 해 오는 과정에서 발생한 돌연변이로 인해 생겨났다. 사람의 눈동자 색깔, 얼굴 형태와 같이 빈번하게 발생하는 것 뿐만 아니라, 때로는 유전 질병처럼 드문하게 발생하기도 한다. 이러한 돌연변이에 의한 질병은 개체들간의 유전정보를 비교하여 어떠한 근거로 유전 질병이 발생하는지 비교 분석이 가능하게 된다. 일반적으로 생물실험실에서 추출되어지는 지노타입은 배수염색체의 정보이기 때문에, 염색체 각 위치에서 대립 형질이 동질접합체(homozygous) 일 경우에는 {A, C, G, T}와 같이 명확한 형태의 정보를 알 수 있지만, 이질접합체(heterozygous)일 경우에는 형질이 섞여있는 모호한 형태로 추출되어진다. 이러한 지노타입 데이터를 단일한 형태의 하플로타입으로 결정해주는 것을 하플로타입 추론 문제라고 한다.

하플로타입을 추론하는 문제는 다음과 같다.
n * k 개의 행렬이 주어진다. n은 지노타입의 개수이고, k는 지노타입의 길이(SNP의 개수)이다. 염색체의 각 SNP 위치에서 대립 형질이 모두 '0' 혹은 모두 '1'로 이루어진 동질접합체이면 지노타입은 '0' 혹은 '1'로 나타나고, 대립 형질이 '0'과 '1' 모두 나타나는 이질접합체이면 지노타입은 '2'로 나타난다. 목적은 n개의 지노타입에서 이질접합위치에 존재하는 대립유전자 쌍들을 0 혹은 1로 결정하는 것이다.

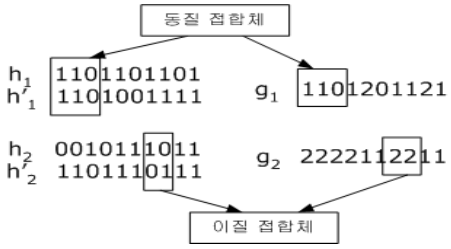

표1. 하플로타입과 지노타입 표현

2. 하플로타입 추론 모델

하플로타입 추론 문제를 해결하기 위해서 접근하는 방법은 4가지로 나눌 수 있다.

첫번째는 Clark[1]이 제안한 추론 알고리즘에 기

반하여 하플로타입을 추론하는 방법이고, 두 번째는 Hawley와 Kidd[2]가 제안한 기대치 최대화(Expectation Maximization) 알고리즘을 이용한 방법이다. 세 번째로 Stephens[3]가 제안한 베이지안 확률 평가와 Gibbs 샘플링을 이용한 확률적 모델에 기반한 방법이 있고, 마지막으로 네 번째는 Gausfield[4]가 제안한 방법으로, perfect phylogeny를 만족하는 하플로타입 집합을 찾는 방법이다.

본 논문에서는 파시모니 모델을 적용하여 하플로타입을 추론하는 방법을 제안한다. 파시모니 모델에서 하플로타입을 추론하는 방법의 목적은 모든 지노타입에서 사용되어지는 하플로타입 패턴 수를 가장 최소화하는 것이다. 파시모니 문제는 Earl Hubell[5]에 의해 제안되어졌고, NP-hard 문제로 알려져있다. 자연 모델에서 발견되는 하플로타입은 가능한 조합의 하플로타입 개수에 비해 매우 작고, 파시모니 모델은 이러한 생물학적 현상을 잘 반영하기 때문에 하플로타입 추론에 부분적으로 적용되어져 왔다.

Gausfield[6]는 하플로타입 추론을 위해 파시모니 모델을 고려한 새로운 방법을 제안하였고, 정수 선형 계획법을 이용하여 하플로타입을 추론하였다.

Wang[7]은 Gausfield가 제안한 파시모니 모델을 적용하여 분기와 한정(branch-and-bound) 기법으로 하플로타입을 추론하였다.

Wang이 구현한 하플로타입 추론 프로그램 HAPAR[7]는 다른 여러 추론 프로그램들(EM-decoder[2], Haplotper, HAPINFERX)보다 더 나은 정확도를 보였고, PHASE[3]와 비교하여 거의 대등한 결과를 얻었다. 하지만, HAPAR는 해결되지 못하는 지노타입이 발생하거나, 지노타입에서 이질접합체의 개수가 많아지면 결과를 도출하는 것이 불가능하게 된다는 문제점을 가지고 있다.

본 논문은 앞서 언급한 파시모니 모델에 기반한 새로운 알고리즘을 제시하고, 파시모니 모델의 다른 연구와 비교 분석한 결과를 보여준다.

3. 하플로타입 추론 알고리즘

모든 지노타입에서 사용되어지는 최소수의 하플로타입 패턴들의 집합을 찾아내는 문제를 고려한다. 이 문제는 NP-hard이므로, 유전자 알고리즘을 이용하여 근사해를 구하는 방법을 설명한다.

3-1. 유전자 알고리즘

유전자 알고리즘[8]은 근사해를 구하기 위해서 생물학적인 진화 모델을 컴퓨터 이론으로 적용한 알고리즘이다. 유전자 알고리즘에서 개체는 가능한 해를 의미하며, 교배, 변이는 더 좋은 해를 탐색하기 위한 유전 연산자로 정의된다. 기본적인 유전자 알고리즘은 다음과 같다.

1. 개체 수 p개만큼 임의의 개체(해)를 생성하여 모집단을 형성한다.
2. 모집단에서 두개의 해 s_1, s_2를 선택하여 교배 연산을 수행하고, 자식 개체 $offspring_i$를 만든다.
3. $offspring_i$ 돌연변이 연산을 수행한다.
4. k개의 자식 개체를 생성할때까지 2부터 반복한다.
5. 우수한 p개의 개체를 선택하여 다음 세대 모집단을 형성한다.
6. 정지조건을 만족할 때까지 2부터 반복 수행한다.

유전자 알고리즘은 다양한 표본들을 얻기위해서 가능한 모든 해집합으로부터 무작위 추출을 통하여 초기 모집단을 구성한다. 전역의 가능한 초기 해로부터 최적 해에 더욱 근접하는 해를 탐색해 나가기 위해 해들이 가지는 우수한 부분들을 유지하여 더 좋은 해를 얻도록 편집한다. 이는 생물들이 자손 번식을 통해 자기와 닮은 개체를 생산하여 환경에 더욱 잘 적응하도록 진화시키는 것과 같은 맥락이다. 다음은 위의 기본적인 유전자 알고리즘을 기반으로 한 구체적인 알고리즘을 기술한다.

① 정의

입력된 지노타입 수를 n, 입력된 개체 수(모집단 크기)는 p, 교배를 통해 생성되어지는 자식 개체 수는o, 반복할 세대 수는 t라고 한다.

모든 지노타입 집합 G는 다음과 같이 정의 한다.

$$G = \{g_i \mid 1 \le i \le n\}$$

각 지노타입 $g_i(1 \le i \le n)$에 존재하는 이질접합체는 0혹은 1로 결정된다. g_i로부터 얻을 수 있는 모든 가능한 하플로타입들의 집합을 H_i라고 할때, 한 원소 h는 g_i를 '커버(cover)한다.'라고 말하고, $h \rightarrow g_i$라고 표현한다.

각 지노타입 g_i를 표현하는 두개의 하플로타입 쌍을 h_i, h_i'라 할때, h_i, h_i'는 'g_i의 결정된 하플로타입 쌍'이라고 말하고, h_i 는 h_i'의 '대립 하플로타입' 이라고 말한다.

각 개체(해) $P_j(1 \le j \le p)$는 다음과 같이 정의 한다.

$$P_j = \{h_i, h_i' \mid 1 \le i \le n\}$$

각 개체 P_j의 크기를 $|P_j|$라 하고, 각 개체 P_j의 하플로타입 원소를 x_k라 할때, P_j는 다음과 같이 표현할 수 있고,

$$P_j = \{x_k \mid 1 \le k \le |P_j|\}$$

각 하플로타입 x_k가 커버하는 지노타입의 집합 C_k은 다음과 같다.

$$C_k = \{g_i \in G \mid x_k \rightarrow g_i\}$$

② 개체 생성

모든 j에서 각 g_i에 대한 난수를 발생키켜 h_i를 얻고, h_i를 이용하여 대립 하플로타입 h_i'을 구한다.

③ 적합도 계산 및 평가

각 개체 P_j의 적합도(fitness) $f(P_j)$는 다음과 같이 정의한다.

$$f(P_j) = 1 \;/\; |P_j|$$

④ 선택

1. 각 개체 P_j가 선택되어질 확률 $pr(P_j)$은 다음과 같이 정의한다.

$$pr(P_j) = f(P_j) \;/\; \sum_{m=1}^{p} f(P_m) \quad (1 \le j \le p)$$

2. 교배를 수행할 개체들은 확률에 의한 선택 메카니즘을 이용하여 결정한다. 이때 선택을 위해서 확률값 $pr(P_j)$을 이용하여 누적 확률 테이블인 룰렛 $R(P_j)$를 다음과 같이 구성한다.

$i = 0$ 일때, $R(P_j) = pr(P_j)$

$i \neq 0$ 일때, $R(P_j) = pr(P_j) + pr(P_{j-1})$

3. 난수 $r(0 \le r < 1)$를 발생시키고, $r < R(P_j)$인 P_j를 선택한다.

⑤ 교배

선택한 두 부모 개체를 각각 P_1, P_2라 한다.

두 부모 개체의 병합 P'는 다음과 같이 정의한다.

$$P' = P_1 \cup P_2$$

1. 각 하플로타입 $x_k \in P' (1 \le k \le |P'|)$가 선택되어질 확률을 구한다.

$$pr(x_k) = |C_k| \;/\; \sum_{m=1}^{|P'|} |C_m| \quad (1 \le k \le |P'|)$$

2. $V = G$, $W = P'$라 하고, 새로운 개체 $P'' = \varnothing$를 만든다.

3. 각 $x_k \in W$의 확률 $pr(x_k)$을 이용한 룰렛 $R(x_k)$를 만들고, x_k를 선택한다.

4. x_k로 커버가능한 지노타입집합 C_k를 구하고, 각 원소 g_i의 두 하플로타입 쌍은 x_k, x_k'으로 한다.

5. $P'' = P'' \cup \{x_k, x_k'\}$, $W = W - \{x_k\}$, $V = V - G$

6. $V = \varnothing$ 일때까지, 3번부터 반복한다.

⑥ 돌연변이

자식개체 P''에 대하여 돌연변이가 일어날 확률 μ에 의한 돌연변이 연산을 수행한다.

$$P'' = M(P'')$$

⑦ 재 평가 및 다음 세대 결정

교배를 통해 생성된 모든 자식개체 $O_j (1 \le j \le o)$와 부모개체 $P_j (1 \le j \le p)$를 병합한 재 평가 개체 집합 S는 다음과 같이 정의한다.

$$S = \{P_i \;,\; O_j \mid 1 \le i \le p,\; 1 \le j \le o\}$$

재 평가 개체 집합에서 적합도 내림차순으로 정렬하여 S'이라 하고, 상위 p개의 개체를 모집단 $P_j (1 \le j \le p)$로 재구성한다.

⑧ 반복 수행

세대수 t만큼 반복적으로 수행한다.

4. 성능 실험

이번 장에서는 파시모니 모델에 기반한 방법들의 실험 결과를 비교 분석한다. 입력되어진 n개의 모든 지노타입에서 각 지노타입의 하플로타입 쌍을 결정할 때, k개의 지노타입이 실제와 일치할 때, 정확도는 k / n가 된다. 파시모니 모델에 기반한 HAPAR[7]는 다른 방법으로 구현되어진 하플로타입 추론 프로그램과 비교해서 비슷한 정확도를 가진다.

그림 1, 2는 분기와 한정 기법으로 구현되어진 HAPAR[7]와 본 논문에서 제안한 추론 방법을 이용하여 구현되어진 GA-Haplotyping의 결과를 비교한 것이다. 각 실험은 지노타입의 샘플 개수에 따라 10개의 다른 테스트 데이터를 이용하여 실험하였으며, 본 논문에서 제안한 GA_Haplotyping은 개체수 100, 세대 수 20, 교배 개체 비율 0.5, 돌연변이 비율 0.01의 일률적인 수행환경을 유지하였다. 수행시간은 지노타입 샘플 개수가 40개 이하이고 SNP 길이가 15 이하인 경우 10-20초가 소요되었고, 지노타입 샘플 개수가 60개 이상이고 SNP 길이가 30 이상인 경우 2-3분 정도가 소요되었다.

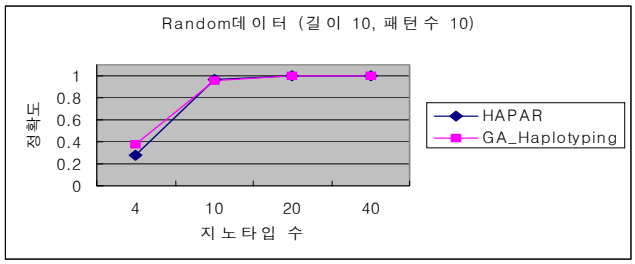

그림 1. 임의 생성 데이터 (길이 10, 패턴수 10)

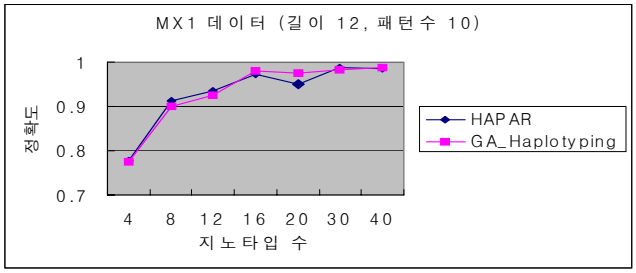

그림 2. 실제 MX1 데이터 (길이 12, 패턴수 10)

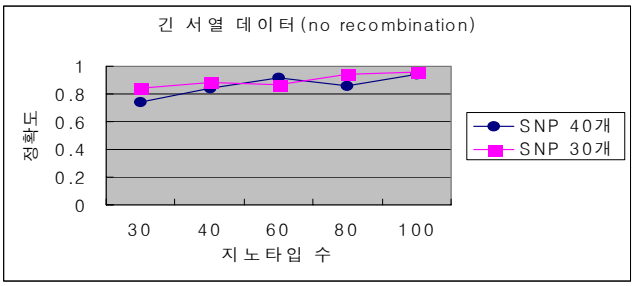

그림 3. 긴 서열 데이터 처리

4-1. 임의 데이터 테스트

임의로 하플로타입 패턴을 랜덤하게 생성한다. 만들어진 패턴들을 조합하여 2*n개 하플로타입 쌍을 만들고, 이로부터 순차적인 n개의 지노타입 샘플 데이터를 생성한다. 그림 1은 지노타입 길이(SNP 개수)가 10개 일때 성능을 비교한 결과이다. 지노타입 샘플 개수가 매우 작을 경우는 두 프로그램 모두 정확한 추론이 불가능 했고, 10개 이상일때, 정확도는 0.95이상으로 상당히 좋은 결과를 얻을 수 있었다.

4-2. 실제 MX1 데이터 테스트

그림 2은 지노타입 길이가 12인 실제의 하플로타입 MX1[9]의 10개의 하플로타입 패턴으로부터 실제로 출현되어진 빈도수를 이용하여 지노타입을 조합하였다. 이 경우 지노타입 샘플 개수가 작은 8~12는 HAPAR가 좀 더 정확한 결과를 나타냈고, 지노타입 샘플 개수가 12~20정도까지는 GA_Haplotyping이 좀 더 나은 결과를 보였으며, 30개 이상부터는 둘다 거의 비슷한 정확도 보였다.

4-3. MS 데이터 생성기를 통한 데이터 테스트

그림 3는 실제의 SNP와 유사한 형태의 매우 긴 하플로타입 서열을 생성하기 위해서 MS 프로그램[10]을 이용하여 지노타입 길이(SNP 개수)가 30개, 40개인 하플로타입 데이터를 얻은 뒤 하플로타입을 추론했다.(재조합 파라미터는 0으로 조정하였다 - 재 조합 발생하지 않음 가정) 같은 지노타입 샘플 개수에서는 SNP 길이가 작은 지노타입 길이 30개를 추론할때, 좀 더 정확한 결과를 얻을 수 있었고, 전체적으로 평균 0.9정도의 정확도를 얻을 수 있었다. HAPAR는 SNP 길이가 대략 20개 이상일 경우 수행시간이 너무 오래걸려 결과를 얻지 못해 비교를 할 수 없었다.

5. 결론

HAPAR는 기존의 다른 프로그램과 비교해서 정확도가 높고, PHASE와 비슷한 정확도를 가진 프로그램이다. 본 논문에서 제안한 방법을 이용하여 구현한 GA-Haplotyping은 HAPAR와 성능을 비교하였을 때, 객관적으로 더 좋은 성능을 보였다. SNP의 길이가 20개 이하일 경우 HAPAR와 비슷한 정확성을 보인다. 그러나, SNP 길이가 20개 이상인 경우 HAPAR는 수행하는 시간이 너무 길어 결과를 얻지 못했지만, GA-haplotyping은 SNP수가 30개 이상 50개 이하 일때도, 최대 3분 이내에 수행이 가능하고, 지노타입 샘플이 너무 작아 분석하기 가 매우 힘든 경우를 제외하고는 평균 0.9 이상의 정확성을 보였다. 실험 결과 파시모니 모델을 이용한 하플로타입 추론은 매우 높은 정확도를 보여주었다.

참고문헌

[1] A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. (Mol. Biol. Evol., 7, pp. 111-122) (1990)

[2] M. Hawley and K. Kidd. HAPLO: a program using the EMalgorithm to estimate the frequencies of multi-site haplotypes. (J. Hered., 86, 409-411) (1995)

[3] M. Stephens, N. Smith and P. Donnelly. A new statistical method for haplotype reconstruction. (Am. J. Hum. Genet., 68,978-989) (2001)

[4] D. Gusfield. Haplotyping as perfect phylogeny: conceptual program framework and efficient solutions. (RECOMB'02 pp. 166-175) (2002)

[5] E. Hubbel. Personal Commuication.

[6] D. Gusfield. Haplotyping by pure parsimony. (UC Davis Computer Science Technical Report CSE-2003-2) (2003)

[7] L. Wang, Y. Xu. Haplotype inference by maximum Parsimony (Bioinfomatics vol. 19 pp. 1773-1780) (2003)

[8] 진강규. 유전자 알고리즘과 그 응용 (교우사) (2000)

[9] L. Jin, P. UNDERHILL, V. DOCTOR, R. DAVIS, P. SHEN, L. CAVALLI-SFORZA, P. OEFNER. Distribution of haplotypes from a chromosome 21 region distinguishes multiple prehistoric human migrations (Proc. Natl. Acad. Sci. USA Vol 96, pp. 3796-3800) (1999)

[10] R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. (Bioinformatics vol 18, pp. 337-338) (2002)