대용량 유전자형 데이터에 대한 LD기반의 일배체형 재구성 시스템*

김상준° 여상수 김성권
중앙대학교 컴퓨터공학부
{jjuns°, ssyeo}@alg.cse.cau.ac.kr, skkim@cau.ac.kr

The LD based Haplotype Reconstruction System for Large scale Genotype dataset

Sang-Jun Kim° Sang-Soo Yeo Sung-Kwon Kim
School of Computer Science & Engineering, Chung-Ang University, Seoul, Korea

요 약

유전자 분석기술의 발전은 지놈 프로젝트(genome project)와 햅맵 프로젝트(hapmap project)를 가능하게 하였으며, 이제는 맞춤형 진단 및 신약 개발 등 실제 사업의 구체화를 가져오게 하였다. 실제 사업에 적용시키기 위해서는 비용 절감의 문제를 해결해야 한다. 그래서 대용량의 유전자형(genotype)데이터를 정확하고 빠르게 일배체형(haplotype)으로 재구성해 줄 수 있는 시스템이 생물 산업 및 제약 산업에서 제기되어 지고 있다. 기존의 연구에서 비록 정확성이 높은 알고리즘들이 개발되어 있지만 기존의 방법들은 계산에 필요한 양이 크기 때문에 대용량 데이터에 대한 처리가 불가능하였다. 우리가 제안하는 시스템은 대용량 데이터를 유동적인 크기로 블록을 분할하여 대용량 데이터 처리 문제를 해결하였다. 또한 나누어진 블록에서 나타나는 모호한 이형접합체(heterozygote)의 위상(phase)의 결정 과정에 LD기반의 블록 분할 방법을 이용함으로써, 추론된 결과의 정확률을 높였다. 구현된 시스템의 성능평가는 ms로 구성한 인공데이터를 사용하여 수행하였다.

1. 서 론

각 개인은 피부색을 비롯하여 눈의 색, 머리카락의 형태, 약물에 대한 반응 등 많은 다양성을 지니고 있다. 지놈 프로젝트와 햅맵 프로젝트를 통해 유전자 분석이 이루어졌고, 각 개인의 유전자에 존재하는 변이중에서 인류의 0.1%이상으로 나타나는 SNP(Single Nucleotide Polymorphism)에 의하여 이러한 다양성이 나타난다는 결과가 나왔다. 그래서 앞으로 맞춤 의학과 신약 개발에서 인간의 다양성을 반영하기 위해 SNP의 모음인 일배체형(haplotype)에 대한 관심이 높아지고 있다.

일배체형 데이터는 유전자형(genotype)데이터로부터 위상(phase)를 구분하여 정렬을 하는 일배체형 재구성(Haplotype Reconstruction)을 거쳐서 구성되어진다. 일배체형 재구성으로 사용되는 기법은 생물학적 접근방식(molecular method)과 계산적 접근방식(computational method)으로 나누어 볼 수 있다. 생물학적 접근방식을 통해 Haplotype Reconstruction을 하는 경우 정확하게 Haplotype을 분석하고 빈도(frequency)가 낮은 일배체형의 발굴이 가능한 장점을 지니지만, 적은 sample에 대해서도 많은 비용과 시간이 소요된다. 앞으로 실제 사업에 적용시키기 위하여 이런 비용을 줄여야 하기에 대안으로 대두되는 것이 계산적 접근방식이다. 계산적 접근방식으로 사용하는 경우 생물학적 접근방식에 비해 적은 시간과 비용이 소요된다는 장점이 있지만, 위상 모호성(phase ambiguity)으로 인해 정확률이 낮다는 문제점이 있다.

본 논문에서 우리는 무관한 sample간에 일배체형을 구성하기 위해 구현한 시스템 MarSelHR을 소개한다. 우리가 구현한 시스템은 기존 시스템보다 많은 SNP 부위에 대해서 여러 sample을 갖고 실험을 할 수 있는 시스템이다. 구현한 시스템의 성능을 분석하기 위해 관련 연구에서 많이 인용하고 있는 PL-EM[1], Haplotyper[2], PHASE[3], HAP[4]을 대상으로 ms프로그램[5]으로 구성한 인공데이터를 사용하여 비교하였다.

2. 알고리즘

2.1 위상의 모호성 관련 연구

염색체로부터 SNP을 구분한 유전자형 데이터는 동질접합(homozygote)형과 이질접합(heterozygote)형으로 구성된 두배수체(diploid)로 되어 있다. 이를 일배체형으로 구성하기 위하여 일배체형 재구성을 한다. 계산적 접근법으로 일배체형 재구성을 할 경우에 이질접합 부위의 수(n)에 따라 2^n가지의 경우가 발생하게 되어, 정확률이 낮아지는 원인이 된다. 그림 1에서 이질접합 부위가 3일 때 가능한 일배체형의 개수는 2^3 즉, 8가지가 발생하는 것을 보여준다.

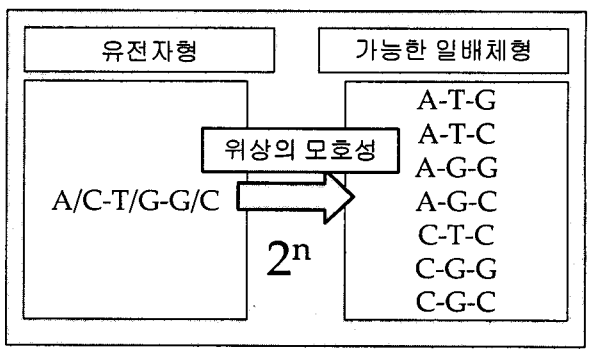

그림 1 위상의 모호성

2.2 시스템 구성

우리가 구현한 시스템은 그림 2에서 보듯이 입력 데이터를 블

* 본 연구는 한국 과학 재단의 기초 과학 연구 사업 과제 (R01-2003-000-11573-0)로 지원받아 수행하였음

록으로 분할 후, 3가지 단계를 거쳐서 일배체형 재구성을 수행한다.

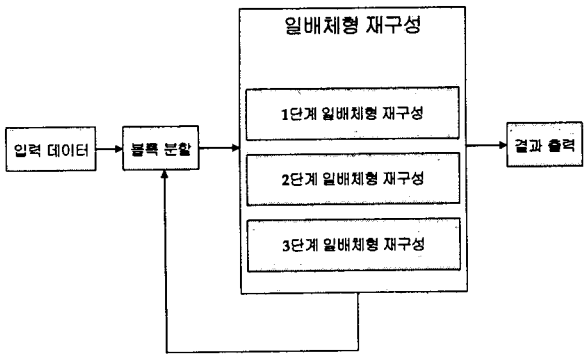

그림 2 시스템 구성도

일배체형 재구성에서 1단계와 2단계는 Clark 알고리즘[6]과 유사하다. 1단계는 동질접합의 유전자형과 하나의 이질접합 부위를 갖는 유전자형에 대해서 참조 테이블을 구성하게 된다. 2단계는 2개 이상의 이질접합 부위를 갖는 유전자형 sample들을 1단계에서 완성된 참조 테이블과 비교하여 모호한 상황이 없다면 해당 유전자형의 결과로 결정하거나 참조 테이블을 구성하게 된다. 3단계는 1단계, 2단계에서 해결하지 못한 모호한 sample들을 LD블록을 이용하여 처리함으로써 하여 모든 sample들에 대하여 처리를 완료하게 된다.

2.3 블록 분할

블록 분할을 하는 것은 대량의 데이터를 일배체형 재구성 할 때 포함되는 이질접합 부위의 수를 줄여서 정확도를 높이기 위함이다. 하지만 블록 단위로 처리를 할 때 생기는 문제는 그림 3과 같이 위상의 모호함이 생길 수 있다는 것이다.

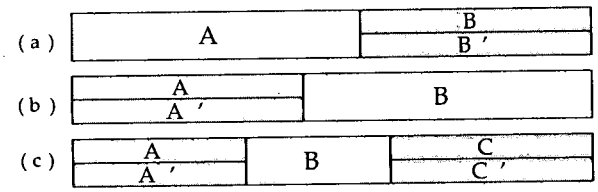

그림 3 블록 단위 처리시에 발생 가능한 위상

그림 3(a)는 동질접합구간의 블록 뒤에 일질접합구간의 블록이 나타나는 경우로 A-B, A-B'로 위상의 모호함은 없다. 그림 3(b)은 이질접합구간의 블록뒤에 동질접합구간의 블록이 오는 경우로 A-B, A'-B로 위상의 모호함은 없다. 하지만 그림 3(c)의 경우는 이질접합구간의 블록 뒤에 동질접합 구간의 블록이 나타나고 그 뒤에 다시 이질접합구간의 블록이 나타나는 경우로 A-B-C, A'-B-C'의 쌍과 A-B-C', A-B-C의 쌍이 생김으로 위상의 모호함이 생기는 것을 보여준다.

그림 3에서 나타나는 위상의 모호함을 해결하기 위해서 그림 4와 같은 블록 범위 알고리즘을 제안한다. 일배체형 재구성이 끝난 구간에서 구간의 끝부분에서 동질접합 구간의 앞에 있는 이질접합 부위를 새로운 블록의 시작점으로 결정을 하고 그 sample에서 동질접합 구간 뒤에 나오는 첫 이질접합 부위를 새로운 블록의 끝부분으로 결정한다.

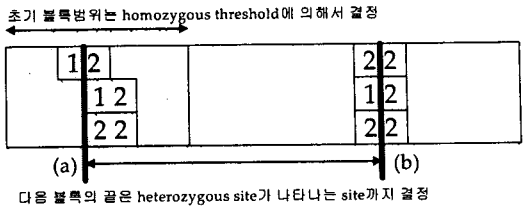

그림 4 블록범위 결정 알고리즘

2.4 LD블록을 이용한 일배체형 재구성

3단계로 구성된 Haplotype Reconstruction에서 1,2단계를 거치고 모호한 genotype sample은 3단계를 거쳐서 최종 haplotype pair를 결정한다. 3단계의 알고리즘은 그림 5와 같다.

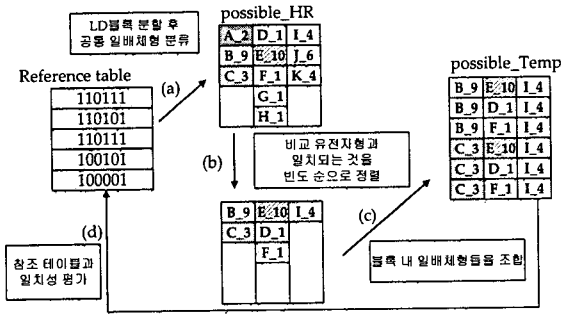

그림 5 LD블록을 이용한 일배체형 재구성 알고리즘

그림 5에서 (a)단계는 참조 테이블에 저장된 일배체형을 대상으로 LD블록을 결정한 뒤 블록단위로 공통 일배체형으로 분류하는 단계이다. LD블록은 Lewontin의 $|D|$[7]을 이용하여 나누었다. (b)단계는 분류된 공통 일배체형을 대상으로 비교 유전자형과 대조하여 동질접합 부위의 일치를 판단하여 빈도순으로 정렬하는 단계이다. (c)단계는 빈도순으로 정렬된 것을 조합하여 가능한 일배체형들을 생성하는 단계이다. (d)단계는 생성된 가능한 일배체형들과 참조 테이블과 비교하여 최종 일배체형 쌍을 선택하게 된다.

3. 실험 환경 및 실험 data

3.1 실험 환경

본 논문에서 구현한 시스템 MarSelHR의 성능평가를 위해 비교한 프로그램은 현재 연구에 많이 인용되어지는 PL-EM V1.0, Haplotyper, PHASE V2.1, HAP V3.0등의 4가지 프로그램이다. 테스트 환경은 MarSelHR은 Pentium4 3.2GHz(HT)의 CPU와 512MB의 메인 메모리를 사용하는 Windows XP시스템을 이용하였고, PL-EM과 Haplotyper, PHASE는 Dual Xeon 550MHz의 CPU와 768MB의 메인 메모리를 사용하는 Linux 시스템을 이용하였으며, HAP은 웹 인터페이스를 사용하였다.

3.2 실험 data

본 논문에서의 실험은 각 프로그램의 일배체형 재구성의 정확률 평가에 목적이 있다. 그래서 ms프로그램을 이용하여 임의 데이터를 생성하여 평가하였다.

ms 프로그램을 이용하여 sample 200개에 대하여 50, 100, 200, 250, 300, 1000개의 크기의 부위에 대하여 생성하였다.

생성된 데이터는 2개의 일배체형에 해당하는 것으로 2개씩 쌍을 이루어 100개의 유전자형으로 조합하여 각 프로그램의 입력 양식에 맞추어 사용하였다.

4. 결 과

4.1 데이터 처리량 과 정확율 비교

데이터 처리량에 대한 관점으로 프로그램들을 비교하였을 때, 표 1에서 보듯이 MarSelHR의 처리량이 1000sites이상으로 가장 높았다.

프로그램명	성공한 부위 수(sites)
PHASE v2.1	100
Haplotyper	250
HAP v3.0	250
PL-EM v1.0	400
MarSel HR	1000

표 1 프로그램별 데이터처리량(100sample기준)

하지만 그림 6에서 보듯이 PHASE와 Haplotyper의 결과가 비교적 높게 나왔고, 우리가 구현한 MarSelHR의 정확률은 다른 프로그램에 비해서 낮은 결과가 나왔다.

Site수	MarSelHR	PL-EM	HAP	Haplotyper	PHASE
50	67	90	51	91	98
100	58	79	52	85	86
200	17	47	11	74	-
250	28	68	33	78	-
300	22	41	-	-	-
400	7	51	-	-	-
500	9	-	-	-	-
1000	13	-	-	-	-

표 2 각 프로그램의 부위 수에 따른 정확률(단위%)

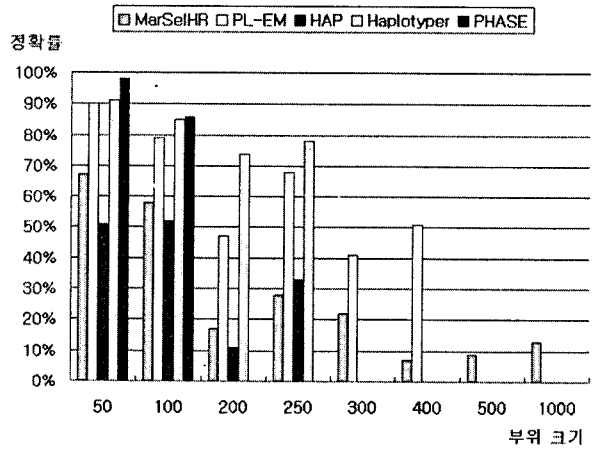

그림 6 프로그램별 데이터처리량 및 정확율 비교(100samples기준)

그래서 100samples,100개의 부위기준으로 실패한 일배체형 재구성의 유전자형 데이터에서 실패한 이질접합 부위의 수를 비교하였다. 그 결과가 그림 7에서 보는 것과 같다.

우리가 구현한 MarSelHR은 다른 프로그램에 비해서 한 부위에서의 실패율이 많아 정확률이 떨어졌음을 알 수 있었다.

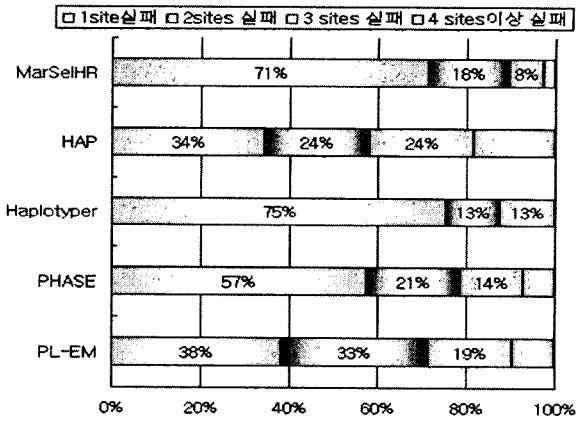

그림 7 프로그램별 실패한 sample의 해당 부위 수 비율(100samples,100부위기준)

5. 결론 및 향후 연구과제

우리는 본 논문을 통해서 대용량 유전자형 데이터를 일배체형 재구성을 할 수 있는 시스템 MarSelHR을 구현하였다. 빠른 처리속도를 위해서 Clark의 알고리즘[6]에 LD기반의 블록분할방법을 적용하여 개선시킨 시스템이다. 다른 연구들에서의 결과는 Clark의 알고리즘은 다른 알고리즘에 비해서 정확률이 매우 낮다고 하였으나 MarSelHR의 성능평가의 내용을 보면 LD블록을 이용하여 다른 프로그램들과 비교할 만한 수준에 이른 것을 알 수 있었다. 또한 MarSelHR이 전체적인 유전자형 sample에 대한 정확률이 낮은 요인은 sample내의 많은 수의 이질접합 부위 중에서 한 부위의 경우에 실패했기 때문이라는 원인을 찾았다.

앞으로 정확률에 대한 보완점과 더불어서 결측치(missing data)가 발생한 유전자형 데이터의 결측치 대치법에 대한 연구를 하도록 하겠다.

6. 참고문헌

[1]Z.S.Qin, T.Niu and J.S. Liu, "Partition-Ligation EM Algorithm for Haplotype Inference with Single Nucleotide Polymorphisms", *Am. J. Hum. Genet.* 71: 1242-47, 2002

[2]Niu, Qin, Xu and Liu, "In silico Haplotype Determination of a Vast Set of Single Nucleotide Polymorphisms.", Technical report, Department of Statistics, Harvard University, 2001

[3]Stephens, M., Smith, N., and Donnelly, P., "A new statistical method for haplotype reconstruction from population data", *American Journal of Human Genetics*, 68, 978-989, 2001.

[4]http://www1.cs.columbia.edu/compbio/hap/data_submission.htm

[5]Hudson,R.R., "Generating samples under a Wright-Fisher neutral model of genetic variation.", *Bioinformatics*, 18:337-338, 2002

[6]A.Clark. "Inference of haplotypes from PCR-amplified samples of diploid populations". *Molecular Biology and Evolution*, 7(2): 111-22,1990.

[7]Lewontin RC, "The interaction of selection and linkage. I. General considerations; heterotic models.", *Genetics* 49: 49-67, 1964