Haplotype Reconstruction 소프트웨어의 성능 평가 및 비교

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

The Performance Evaluation and Comparison of Softwares for Haplotype Reconstruction

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

요 약

SNP(Single Nucleotide Polymorphism)은 생물학적 다양성에 관한 연관성 연구(Association Study)에서 이용되어지고 있다. haplotype을 구하기 위해 genotype data를 Haplotype Reconstruction을 하여 한 가닥씩 분리를 한다. Haplotype Reconstruction의 방법은 생물학적 접근법(molecular method)과, 계산적 접근방법(*in-silico* method)으로 연구되고 있다. 계산적 접근법은 생물학적 접근법에 비해 적은 비용과 시간이 소요되는 장점을 지니지만, phase problem으로 인하여 생물학적 접근법에 비해 정확도가 낮다는 단점을 갖는다. 이런 문제를 해결하기 위한 여러 알고리즘들과 프로그램들이 연구 및 개발되고 있다. 본 논문에서는 현재 개발된 프로그램들에 대해서 다양한 테스트를 통한 각 프로그램의 성능 비교를 하였고, 특성과 문제점을 파악하였다.

1. 서 론

각 개인은 피부의 색을 비롯하여 눈의 색, 머리카락의 형태, 약물에 대한 반응등에 많은 다양성을 지니고 있다. 인간의 유전체 안에는 이러한 다양성에 영향을 주는 SNP(Single Nucleotide Polymorphism)이 대략 3백만 개 정도 존재한다고 알려지고 있다. 한 염색체 안에서 인접한 SNP만을 연결한 것을 haplotype이라고 한다. 개별 SNP보다 haplotype으로 얻는 정보가 더 효율적이다 하여 haplotype에 많은 관심을 갖게 되었다[1].

Diploid로 구성된 인간의 염색체에서 SNP을 찾아내기 위해 Genotyping을 한다. 이때 생성된 diploid data에서 SNP을 일렬로 구분하여 정렬하는 과정을 Haplotype Reconstruction이라 한다. 이러한 방법은 크게 생물학적 접근법(molecular method)과 계산적 접근법(*in-silico* method)으로 나눌 수 있다. 생물학적 접근법을 통해 Haplotype Reconstruction을 하는 경우 정확하게 Haplotype을 분석하고 빈도(frequency)가 낮은 Haplotype의 발굴이 가능한 장점을 지닌다. 계산적 접근법으로 시도하는 경우 생물학적 접근법에 비해 적은 시간과 비용이 소요되는 장점이 있지만, phase ambiguity(모호성 문제)로 인해 정확율이 낮다는 문제점이 있다.

우리는 무관한 sample간에 haplotype을 구하기 위해 계산적 접근법에서 phase ambiguity를 해결을 목표로 삼고 있다. 이를 위해 관련 연구에서 많이 인용하고 있는 PL-EM[2], Haplotyper[3], PHASE[4], HAP[5]을 대상으로 비교하여 보았다.

2. 알고리즘

2.1 Phase problem관련 연구

염색체로부터 SNP을 구분한 genotype data는 homozygous (동질접합)형과 heterozygous(이질접합)형으로 구성된 diploid로 되어 있다. 이를 haplotype으로 구성하기 위하여 haplotype reconstruction을 한다. 계산적 접근법으로 haplotype reconstruction을 할 경우에 heterozygous인 SNP수(n)에 따라 2^n가지의 경우가 발생하게 되어 정확율을 낮게 하는 원인이 된다. 그림1에서 heterozygous인 SNP이 3가지 일 때 가능한 haplotype의 개수는 2^3 즉, 8가지가 발생하는 것을 보여준다.

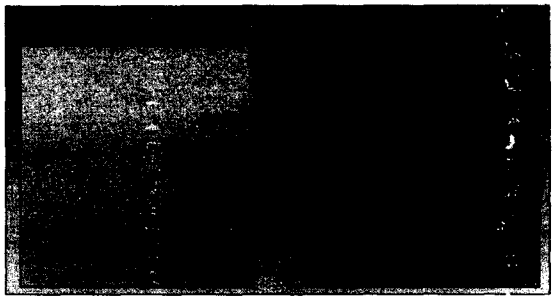

그림 1 Phase problem

2.2 실험 설계

실험할 프로그램은 genotype data를 입력받는다. 입력 data의 sample로부터 추론을 하여 haplotype의 빈도(frequency)의 순서로 출력을 한다. 본 실험은 생물학적 접근법으로 얻은 haplotype data로부터 genotype data로 변형을 하여 각 프로그램에 입력한다. 각각 나온 결과를 원래의 haplotype data와 비교를 하여 추론의 정확율을 계산하고 프로그램별로 비교를 하는 것이다.

실험 대상의 프로그램 중에 PL-EM과 Haplotype의 경우에는 SNP길이와 sample개수에 대한 제한이 있어서 Daly data[6]의 경우에는 한번에 프로그램을 실행시킬 수가 없었다. 그래서 HAP을 먼저 테스트하고, 그 결과로 나온 블록으로 Daly data를 나누어 PL-EM과 Haplotype과 PHASE를 테스트 하였다.

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

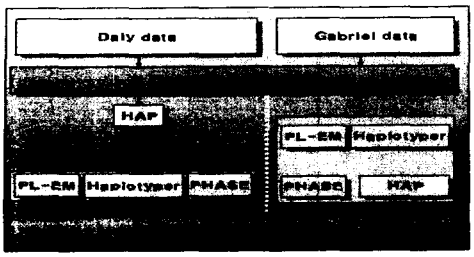

그림 2 테스트 순서 설계

2.3 정확율 비교 프로그램 구현

비교할 프로그램의 추정 결과 데이터와 실제 데이터가 어느 정도 일치하는지 판단하기 위해 분석프로그램을 구현하였다.

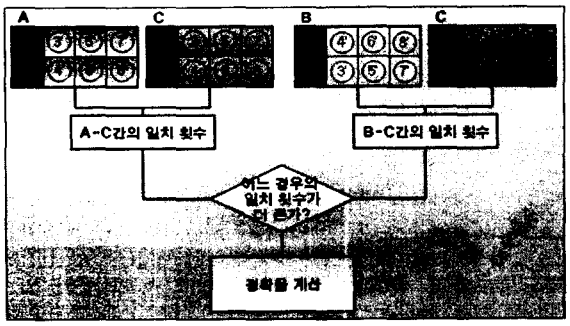

그림 3 테스트 프로그램 추론결과의 정확율 계산 알고리즘 개념도
A: 추론결과 B: 추론결과를 위아래로 바꾼 경우 C: 비교 데이터

비교 대상의 4가지 프로그램은 각 sample단위로 2줄로 haplotype을 출력한다. 그림3에서 A와 B는 sample 1개에 대한 haplotype이고, 서로는 haplotype의 열이 바뀐 경우를 의미한다. C는 genotype으로 변형시켰던 원래의 data이다. A와 B는 C에 대하여 ①과 ②를 동시에 일치할 때 1점씩 부여를 한다. 모든 SNP에 대해 일치성을 판단한 뒤 A, B의 경우 중 많은 점수를 획득한 것에 기준상아 획득한 점수를 SNP의 수로 나누어 주어 sample 하나에 대한 정확율을 구한다. 모든 sample에 대해서 구한 정확율을 평균을 구하여 프로그램의 정확율로 인정한다. 이것은 식으로 표현하면 다음과 같다.

P_{AC} : A와 C의 일치한 점수

P_{BC} : B와 C의 일치한 점수

N_{SNP} : SNP의 수

N_{sample} : sample의 수

$$Accuracy \quad rate = \frac{\sum_{i=1}^{N_{sample}} \frac{max(P_{A,C_i}, P_{B,C_i})}{N_{SNP}}}{N_{sample}}$$

3. 실험 환경 및 실험 data

3.1 실험 환경

본 논문에서의 프로그램 비교의 대상은 현재 연구에 많이 인용되어지는 PL-EM V1.0, Haplotyper, PHASE V2.0.2, HAP V2.0의 4가지 프로그램이다. 테스트 환경은 PL-EM과 Haplotyper, PHASE는 Dual Xeon 550MHz의 CPU와 768MB의 메인메모리를 사용하는 Linux 시스템을 이용하였고, HAP은 웹 인터페이스를 사용하였다.

3.2 실험 data

실험에 사용한 data는 Daly data[6]와 Gabriel data[7]를 사용하였다. Daly data는 아버지, 어머니, 자식의 3명으로 구성된 유럽인 129가정의 가계 정보를 포함하여 129 X 3으로 총 378명의 SNP정보로 구성되어 있다.

Gabriel data는 크게 4가지로 구분되어진다. European의 12가족의 93명의 가계도정보와 SNP정보로 이루어진 popA와 50명의 무관한 아프리카계의 미국인의 SNP정보인 popB, 42명의 무관한 Japanese와 chinese의 SNP정보인 popC, 그리고 Yoruba족 32가족의 63인의 가계도 정보와 SNP정보인 popD로 구성되어 있다.

이 중에 사용할 데이터의 부분은 Daly데이터의 경우 가계도 정보를 제외하고 자식부분만을, Gabriel데이터의 경우 50명의 아프리카계의 미국인의 정보인 popB부분을 사용하였다.

4. 결 과

2절에서 소개한 분석프로그램으로 4가지 프로그램에 대한 결과는 아래 표1과 같다.

표 1 정확율 분석 프로그램에 의한 프로그램별 결과

Block	SNP수	PL-EM	Haplotyper	PHASE	HAP
7b	13	[illegible]	0.8138	[illegible]	0.8015
37a	47	[illegible]	0.8455	[illegible]	0.8396
7a	55	[illegible]	0.8415	[illegible]	0.8353
39a	64	[illegible]	0.8438	[illegible]	0.8134
19a	72	[illegible]	0.8672	[illegible]	0.8331
40a	80	[illegible]	0.8600	[illegible]	0.8160
24a	92	[illegible]	0.8172	0.7609	0.7963
block2	9	[illegible]	0.8587	[illegible]	0.8691
block4	8	[illegible]	0.9041	0.9230	0.9215
block7	5	[illegible]	0.8946	0.8372	0.8372
block11	6	[illegible]	0.8023	0.7981	0.7907
block12	3	[illegible]	0.8269	0.8320	0.8320
block13	7	[illegible]	0.8594	0.8680	0.8549

표 1의 내용을 바탕으로 sample수와 SNP수가 일정할 때, missing data의 입력에 따른 정확율을 그림 4와 5에서 보이고 있다. sample수가 많을수록 입력data의 missing data가 많이 포함됨을 볼 수 있었다. sample수가 많음에 따라 정확율은 높지만, missing data의 수에 대해서 정확율의 차이가 크게 나타남을 볼 수 있었다.

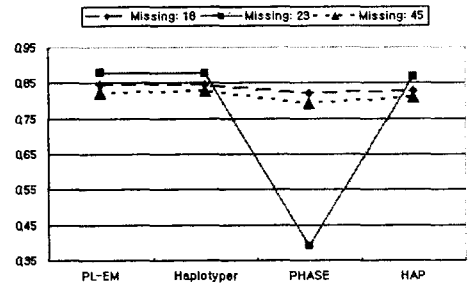

그림 4 sample수 50일 때 missing data에 따른 정확율

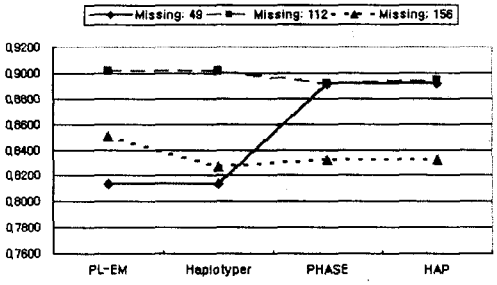

그림 5 sample수 129일 때 missing data에 따른 정확율

일정한 sample수일 경우에 SNP수에 따른 정확율을 그림 6과 7에서 보이고 있다. 대체로 SNP수의 변화에 따른 정확율의 변동 량이 매우 미비했다. 하지만 missing data의 수에 대해 눈에 띄는 변동 량이 간혹 발견되었는데 이 경우의 변동 량의 폭은 sample수에 따라 영향을 받고 있음을 볼 수 있었다.

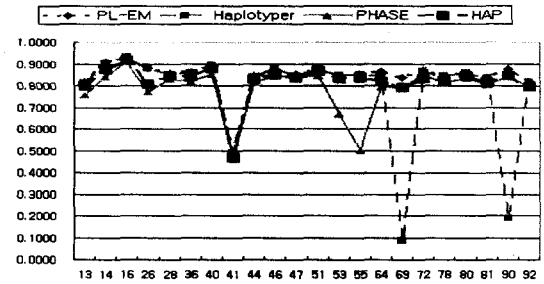

그림 6 sample 50일 때 SNP수에 따른 정확율

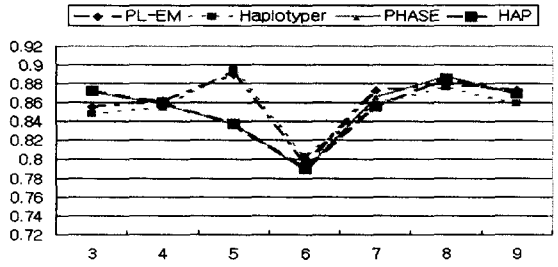

그림 7 sample수 129일 때 SNP수에 따른 정확율

100개 이상의 SNP에 대해서 sample수에 따른 정확율을 그림 8과 9에서 보이고 있다. PL-EM과 Haplotyper의 경우 SNP수와 sample수에 따른 제한이 있어서 129sample의 103SNP인 Daly data의 경우 테스트 하지 못하였다. sample수가 많고 SNP수가 많으면 비교해야 하는 총 SNP수가 매우 커지게 된다. 추론해야할 총 SNP수가 커질수록 그 안에 포함되어 있는 missing data로 인해 정확율의 차이가 큼을 볼 수 있었다.
처리 시간은 프로그램별로 입력data의 양에 따라 비례적으로 증가하였다. PHASE의 경우는 Daly data를 전체를 테스트하는데 걸린 시간이 약26시간 걸린데 비해, HAP의 경우 10분정도의 비교적 짧은 시간이 소요되는 결과를 보여 프로그램 수행시간에서 많은 차이를 보였다.

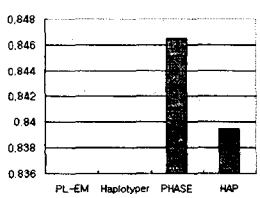

그림 8 129개의 sample수와 103개의 SNP인 Daly data결과

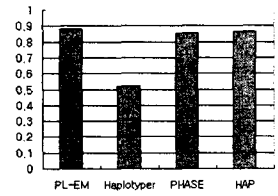

그림 9 50개의 sample수와 Gabriel data 41a

5. 결론 및 향후 연구과제

본 논문의 실험을 통하여 비교 대상으로 사용했던 프로그램들은 입력받은 sample수와 SNP수, 그리고 입력 data에 포함되는 missing data로 인해 정확율에 영향을 받았다. sample수가 많을수록 SNP수에 따른 정확율의 변동 량은 적었다. 하지만, 정확율은 sample수가 많은 경우가 적은 경우보다 낮은 결과를 보였다. 또한 입력 data에서의 missing data양이 많아질수록 정확율은 낮았다. 또한 현재까지의 Haplotype Reconstruction 프로그램들은 대부분 79~86%정도의 범위 안에서 정확율을 유지하고 있음을 보았다.

이 정도의 정확율은 아직 전산 통계적 접근 방법으로 생물학적 실험법을 대체할 정도의 수준은 아니며, 발전의 가능성을 가지고 있다고 말할 수 있다. 대용량의 SNP처리를 위하여 처리 속도를 빠르게 해야 하는 문제와 missing data에 대한 처리를 해결하는 방법을 모색하여야겠다. 이번 실험의 결과를 토대로 더 나은 알고리즘과 프로그램 구현을 시도하도록 하겠다.

6. 참고문헌

[1]Xing Wang, "HIT: a Haplotype Inference Testbed", *CAPSL*, 2003

[2]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

[3]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

[4]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.

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

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

[7]Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D. "The Structure of Haplotype Blocks in the Human Genome", *Science*, 296(5576):2225-9, 2002