Haplotype Reconstruction 소프트웨어의 성능 평가 및 비교

김성준, 나경학, 이상수, 김성현
중앙대학교 컴퓨터공학부
{jinse, bike, ssye}@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 이용되며, haplotype를 구하기 위해 genotype data를 Haplotype Reconstruction을 하여 가정된 통계적 분석을 한다. Haplotype Reconstruction의 방법은 생물학적 접근법(molecular method)과, 계산적 접근법(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 단점이 있다. 이런 문제를 해결하기 위한 여러 알고리즘들과 프로그램들이 연구 및 개발되고 있다. 본 논문에서는 현재 개발된 프로그램들에 대해서 다양한 테스트를 통한 각 프로그램의 성능을 비교하였고, 성능과 문제점을 파악하였다.

1. 서론

각 개인은 피부의 색을 비롯하여 눈의 색, 머리카락의 형태, 악물에 대한 반응등에 많은 다양성을 가지고 있다. 인간의 유전체 안에는 이러한 다양성에 대해 SNP(Single Nucleotide Polymorphism)이 대략 3백만 개 정도 존재한다고 알려져 있다. 한 양체 안에서 인접한 SNP만을 연결한 것을 haplotype이라고 한다. 개별 SNP보다 haplotype로 얻는 정보가 더 효율적이다. 하여 haplotype에 많은 관심을 갖게 되었다. 

Diploid로 구성된 인간의 염색체에서 SNP를 찾아내기 위해 Genotyping을 한다. 이때 생성된 diploid data에서 SNP를 일렬로 구분하여 정밀하게 관찰하는 과정을 Haplotype Reconstruction이라고 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 Haplotype를 분석하고 변도(frequency)가 낮은 haplotype의 발견이 가능한 장점을 지니고 있다. 계산적 접근법으로 시도하는 경우 생물학적 접근법에 비해 적은 시간과 비용이 소요되는 장점이 있지만, phase ambiguity(모호성 문제)로 인해 정확도가 낮다는 문제점이 있다.


2. 알고리즘

2.1 Phase problem 관련 연구

염색체로부터 SNP를 구분한 genotype data는 homozygous (동질질합)형과 heterozygous(이질질합)형으로 구성된 diploid로 되어 있다. 이를 haplotype로 구성하기 위하여 haplotype reconstruction을 한다. 계산적 접근법으로 haplotype reconstruction을 할 경우에 heterozygous인 SNP수(n)에 따라 2가지의 경우가 발생하게 되어 정확도를 낮게 하는 원인이 된다. 그러므로 heterozygous의 경우는 2개가 발생하는 경우에 3가지일 때 가능한 haplotype의 개수는 2^3 즉, 8가지가 발생하는 경우를 보여준다.

그림 1 Phase problem

2.2 실험 설계

실험할 프로그램은 genotype data를 입력받는다. 입력 data의 sample로부터 추론을 하여 haplotype의 변도(frequency)의 순서로 출력한다. 본 실험은 생물학적 접근법으로 얻은 haplotype data로부터 genotype data로 변환을 하여 각 프로그램에 입력한다. 각각의 결과를 본래의 haplotype data와 비교하여 추론의 정확도를 계산하고 프로그램별로 비교하는 것이다.

2.3 정확도 비교 프로그램 구현

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

이런 작업을 수행하던 중 필자는 박현주관에 전혀 비슷한 일에 임직원이 개발한 Haplotyper 프로그램에 대해 근근히 비교하였다. 앞서 언급한 것과 같이, 이는 실제 데이터와 비교하여 결과를 추정하는 프로그램이었기 때문이다.

이 프로그램을 이용하여 결과를 추정하는 방법은 다음과 같다. 먼저, Haplotyper의 결과를 기준으로 실제 데이터와 비교하여 정확도를 평가한다. 이때, 정확도는 실제 데이터와 비교하여 정확하게 일치하는 부분의 비율로 정의된다.

이 프로그램은 정확도를 계산하는 데 사용되었으며, 이는 Haplotyper의 정확도를 추정하는 데 도움이 된다. 이를 통해 프로그램의 성능을 평가할 수 있으며, 이를 통해 프로그램의 향상이 가능하다.

3. 실험 환경 및 실험 데이터
3.1 실험 환경

본 논문에서는 PL-EM V1.0, Haplotyper, PHASE V2.0와 HAP V2.0의 4가지 프로그램을 테스트하였다. 이는 PL-EM과 Haplotyper, PHASE는 Dual Xeon 550MHz의 CPU와 768MB의 메모리 모듈을 사용하는 Linux 시스템을 사용하였다. HAP은

원 인터페이스를 사용하였다.

3.2 실험 데이터


표 1 정확도 분석 프로그램에 의한 실험 결과

<table>
<thead>
<tr>
<th>Block</th>
<th>SNP 수</th>
<th>PL-EM Haplotyper PHASE HAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>7b</td>
<td>13</td>
<td>0.8138</td>
</tr>
<tr>
<td>37a</td>
<td>47</td>
<td>0.8455</td>
</tr>
<tr>
<td>7a</td>
<td>55</td>
<td>0.8415</td>
</tr>
<tr>
<td>36a</td>
<td>64</td>
<td>0.8438</td>
</tr>
<tr>
<td>36a</td>
<td>64</td>
<td>0.8672</td>
</tr>
<tr>
<td>40a</td>
<td>80</td>
<td>0.8600</td>
</tr>
<tr>
<td>24a</td>
<td>92</td>
<td>0.8172</td>
</tr>
<tr>
<td>Block2</td>
<td>9</td>
<td>0.8587</td>
</tr>
<tr>
<td>Block4</td>
<td>8</td>
<td>0.9041</td>
</tr>
<tr>
<td>Block7</td>
<td>5</td>
<td>0.8946</td>
</tr>
<tr>
<td>Block11</td>
<td>6</td>
<td>0.8023</td>
</tr>
<tr>
<td>Block12</td>
<td>3</td>
<td>0.8269</td>
</tr>
<tr>
<td>Block13</td>
<td>7</td>
<td>0.8594</td>
</tr>
</tbody>
</table>

표 1의 내용을 바탕으로 sample 수와 SNP 수가 일정할 때, missing data의 입력에 따른 정확도를 평가한 결과는 다음과 같다. sample 수가 많을수록 입력 data의 missing data가 많이 포함될 수 있으며, sample 수가 많을수록 정확도는 높지만, missing data의 수에 따라서 정확도의 차이가 크게 나타날 수 있었다.
5. 결론 및 후속 연구개체

본 논문의 실험을 통하여 비교 대상으로 사용했던 프로그램들은
방법론은 sample수와 SNP수, 그리고 입력 데이터에 포함된
missing data로 인해 정확도에 영향을 받았다. sample수가
많을수록 SNP수에 따른 정확도의 변동량은 적었다. 하지만,
정확도는 sample수가 적은 경우가 적은 경우보다 낮은 결과를
얻었다. 또한 입력 데이터에서의 missing data량이 많아지면
정확도는 낮아졌다. 또한 현재까지의 HapMap Reconstruction
프로그램들은 대부분 79~86%정도의 범위 안에서 정확도를 유
지하고 있음을 보였다.

이 결과는 정확도는 아직 전산 통계적 접근 방법으로 생물학
적 실험법을 대체할 정도의 수준은 아니며, 발전의 가능성을
가지고 있다고 말할 수 있다. 대응형의 SNP처리를 위하여 처
리 속도를 빠르게 해야 하는 문제와 missing data에 대한 처리를
해결하는 방법을 모색하여야 한다. 이 산 실험의 결과를 토대
로 더욱 나은 알고리즘과 프로그램 구현을 시도하도록 하겠다.

6. 참고문헌