품질정보의 사용유무에 따른 하플로타입 페이징의 결과 차이

이종찬, 나중채†
세종대학교 컴퓨터공학과
e-mail : jchlee91@gmail.com, jcna@sejong.ac.kr

Difference in Haplotype Phasing According to the Use of Quality Information

Jong-Chan Lee, Joong Chae Na
Dept. of Computer Science & Engineering, Sejong University

요 약

인간 유전자의 SNP 서열 정보를 통해 하플로타입을 추정하는 하플로타입 페이징은 생명공학분야에서 중요한 연구분야이다. 최근에는 SNP 데이터가 많아짐에 따라 많은 하플로타입 페이징 알고리즘들이 제시되었다. 본 논문에서는 SNP 데이터의 오류로 인한 하플로타입 페이징의 한계점과 이를 해결하기 위한 품질정보의 사용에 관한 문제점을 언급한 후 이와 관련된 실험을 통해 품질정보가 하플로타입 페이징의 결과에 미치는 영향을 알아본다. 실험은 기존의 하플로타입 페이징 알고리즘을 사용하여 품질정보의 사용 유무에 따라 하플로타입 페이징 결과를 비교하는 과정으로 진행되었다. 실험 결과 하플로타입 페이징에 과정에서 품질정보를 사용하는 것은 품질정보를 사용하지 않았을 때 보다 좋은 결과를 보여주었다.

1. 서론

인간은 DNA로 이루어진 23쌍의 염색체를 가지고 있으며, 이들 서열의 정보는 유전체 분석 연구의 기본이 되는 중요한 정보이다. 최근에는 NGS(Next Generation Sequencing)와 같은 서열 결정을 대용량으로 쉽게 할 수 있는 방법들이 개발됨으로 인해, 다룰 수 있는 서열 정보의 양이 늘어났다. DNA 서열 기반 변이 연구 중 핵심적인 분야 중 하나가 인간 유전체에서 가장 빈번하게 발생하는 단일 염기 다형성 SNP(Single Nucleotide Polymorphism)이다.

SNP는 DNA 염기 서열 중 같은 종의 서로 다른 객체에서 나타난 한 염기에서 나타나는 차이를 의미한다[1]. 하플로타입(haplotype)은 동일 염색체상 여러 위치에서의 대립형질의 조합을 의미한다. SNP들의 조합으로 구성된 SNP matrix가 주어졌을 때, 이로부터 한 쌍의 하플로타입을 만들어내는 하플로타입 페이징(haplotype phasing)은 아주 중요한 연구분야이다.

하지만 하플로타입에 페이징에 사용되는 SNP matrix에는 기본적으로 오류가 있는 값들이 존재한다. 이는 하플로타입 페이징에 어려움을 주며 이를 처리하기 위한 효율적인 방법을 제시하는 것이 하플로타입 페이징 알고리즘을 설계하는데 있어 중요한 문제이다. SNP matrix에는 각 요소마다 품질정보가 존재하는데 이는 NGS 과정에서의 base calling 에러율과 관련된 정보를 가지고 있다. 이 품질정보는 SNP matrix의 각 요소들이 얼마나 정확한 정보인지 나타내 주는 값이며 이를 이용하여 하플로타입 페이징을 하는 것이 일반적이다.

현재까지 많은 하플로타입 페이징 알고리즘들이 품질정보를 사용하여 문제를 해결하였다. 하플로타입 페이징 알고리즘중 하나인 DBM[3]의 경우 품질정보를 사용한 markov 모델을 이용하여 하플로타입 페이징을 진행한다. 또한 ProbHap[4]의 경우 품질정보를 이용한 확률적모델을 사용하여 하플로타입 페이징을 한다. 이 외에도 유전알고리즘, 그리디, 확률론적 접근을 이용한 여러 알고리즘들이 존재하며 이들 대부분 품질 정보를 사용한다.

하지만 SNP matrix에서 오류가 있는 부분의 품질이 오류가 없는 부분보다 높은 품질로 나타나는 경우가 존재한다. 이는 품질 정보를 사용하는 하플로타입 페이징에서 문제가 될 가능성이 있다. 따라서 본 논문에서는 품질 정보의 사용 유무에 따른 하플로타입 페이징 결과의 정확도를 비교하기 위해 기존에 존재하는 하플로타입 페이징 알고리즘을 사용하여 실험을 진행한다. 이를 통해 품질 정보가 하플로타입 페이징의 결과에 어떤 영향을 미치는지 알아본다.

---

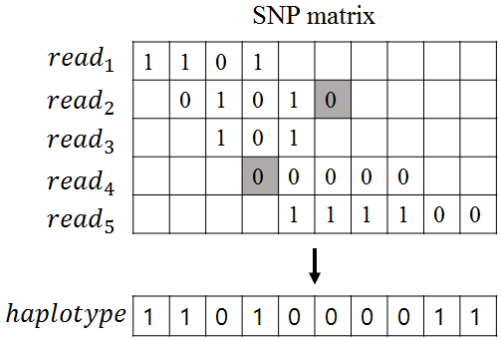
† 교신저자.

이 논문은 미래창조과학부 및 정보통신기술진흥센터의 SW중심대학 지원사업(R7718-16-1005)과 2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(2014R1A1A1004901).

2. 품질정보

하플로타입 페이징이란 SNP matrix로부터 한 쌍의 하플로타입을 결정하는 문제이다(그림 1). SNP matrix는 2차원 matrix로 구성되어 있으며 각 요소는 '1', '0'으로 표현된다. SNP의 특정 위치의 염기가 같은 위치의 표준염기서열의 염기와 같은 값을 가지는 경우 matrix내의 해당 요소는 '1'로 표현하고, 다른 값을 가지는 경우는 '0'으로 표현한다. SNP matrix의 각 read(행)은 부모 중 한쪽에서 온 것이며, 부모는 서로 보수관계이다. 이런 SNP matrix의 read들 정보를 조합해 한 쌍(부/모)의 하플로타입을 만들어내는 것이 하플로타입 페이징의 기본적인 목표이다.

서론에서 언급했듯이 SNP matrix에서는 오류가 있는 자리가 존재한다. (그림 1)에서 음영으로 처리된 부분들은 오류가 있는 자리를 의미한다. $read_2$, $read_3$, $read_5$는 $read_1$, $read_4$ 와 서로 보수관계이며 하플로타입과도 보수관계인 read들이다. 하플로타입과 대조하였을 때 $read_2$는 01011이 되어야 하플로타입과 완전한 보수관계가 되는 오류가 없는 read가 되고, $read_4$는 10000이 되어야 하플로타입과 완전히 일치하는 오류가 없는 read가 된다. 따라서 $read_2$와 $read_4$의 음영처리된 부분은 오류가 발생한 요소이며 각각 보수가 되어야 올바른 값이다.



(그림 1) SNP matrix를 이용한 하플로타입 페이징

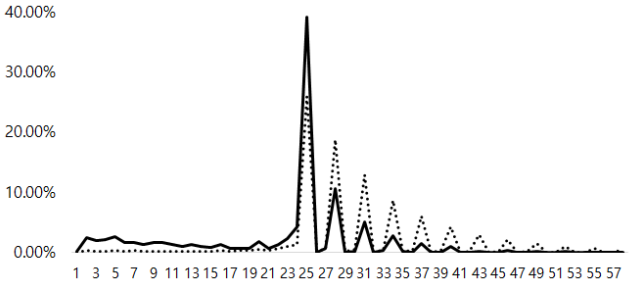
이런 오류들이 없다면 간단한 방법으로 정확한 하플로타입을 만들어낼 수 있지만, 대부분의 SNP matrix에는 이와 같은 오류들이 존재한다. 따라서 문제를 해결하기 위해 모든 가능한 하플로타입을 조합해봐야 하기 때문에 하플로타입 페이징 문제는 NP-hard문제로 남아 있다[9]. 하플로타입 페이징 알고리즘의 설계에 있어 중요한 고려사항 중 하나가 이런 오류를 어떻게 처리할 것인지의 문제다. SNP matrix에서의 품질정보는 오류가 있는 자리에 대한 정보를 가지고 있는 SNP matrix에서의 유일한 정보이기 때문에 오류가 있는 부분을 처리하기 위해서는 해당 부분의 품질 정보를 이용해야 한다.

품질정보는 SNP matrix의 각 요소별로 존재한다. 이 품질정보는 NGS 과정에서 miss call된 확률과 관련된 정보를 가지고 있으며, 이는 SNP matrix 내의 해당 요소가 얼마나 정확한 정보를 가지고 있는지를 의미한다. SNP matirx 내의 품질정보는 phred quality score(그림 2)형태로 표현되어 있다. 이는 base calling 에러 확률에 대한 정보를 0부터 90범위의 정수 형태로 표현한다. 임의의 요소의 품질값이 클수록 에러율이 낮고 해당 요소가 정확함을 의미한다.

$$Q = -10\log_{10}P$$
$$Q: phred\ 품질\ 정보$$
$$P: base-calling\ 에러율$$

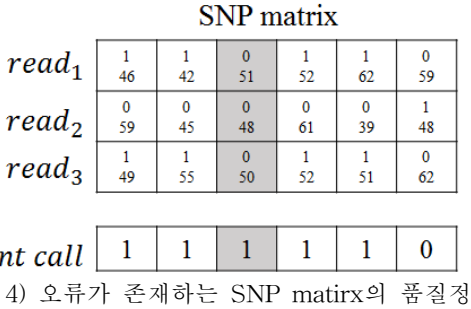
(그림 2) Phred quality score로 표현되는 품질정보

(그림 3)은 SNP matrix에서 오류가 있는 부분의 품질과 오류가 없는 부분의 품질의 분포를 보여주고 있다. 가로축은 품질 정보를 의미하고 세로축은 해당 품질이 SNP matrix에 존재하는 비율을 의미한다. 오류가 존재하는 부분의 품질(실선)은 오류가 없는 부분의 품질(점선)보다 전체적으로 수치가 낮은 분포를 보여주고 있다. 하지만 품질분포가 명확하게 나눠져 있지는 않은 것을 알 수 있다.



(그림 3) 오류가 있는 부분과 없는 부분의 품질 분포

이와 같이 품질 분포가 명확하지 않기 때문에 SNP matrix 내에 (그림 4)와 같은 경우가 존재한다. (그림 4)의 경우 SNP matrix의 read들과 정답(variant call)을 대조해보면 오류가 있는 부분의 품질이 오류가 없는 부분의 품질보다 높은 품질값을 가지는 경우를 보여준다. $read_1$과 $read_3$의 3번째 열에서 오류가 발생하였지만 오류가 발생하지 않은 $read_2$의 품질(48)이 $read_1$(51)과 $read_3$(50)보다 낮은 것을 확인할 수 있다.



(그림 4) 오류가 존재하는 SNP matirx의 품질정보

이와 같이 SNP matrix내의 정확하지 않은 품질정보로 인해 품질정보를 사용하는 하플로타입 페이징의 결과에 문제가 될 가능성이 있다. 따라서 본 논문에서는 품질 정보를 사용하여 하플로타입 페이징을 하는 것이 적합한지에 대한 문제를 이와 관련된 실험을 통해 논의한다.

3 실험 및 결과

본 연구에서 진행하는 실험은 하플로타입 페이징 과정에서의 품질 정보의 사용 유무에 따른 페이징 결과를 비교하여 품질 정보가 하플로타입 페이징 결과에 어떤 영향을 미치는지 알아본다. 실험에 사용되는 하플로타입 페이징 알고리즘은 최근에 개발된 성능이 우수한 ProbHap과 본 저자가 이전 연구에서 제안한 베이지안 네트워크를 이용한 BOAhap[5]이다.

본 실험에서 하플로타입 페이징 결과의 정확도를 측정하기 위한 방법으로는 하플로타입 페이징 알고리즘의 성능을 판단하는 척도 중 하나인 switch error 비율을 사용한다[6]. Switch error는 한 하플로타입의 일정 위치에서 정답과 대조되는 보수가 발생되는 경우를 의미한다. 예를 들어 정답이 11111111이고 하플로타입이 11111110이면, 8번째 위치의 값이 정답과 보수이므로 8번째 위치에서 switch error가 발생하게 된다.

실험에 사용되는 데이터는 가장 대중적인 NA12878 chromosome[7]를 사용한다. NA12878은 22개의 각 chromosome으로 구성되어 있으며 본 실험에서는 22번째 chromosome을 사용한다. 또한 정확도를 판단하기 위해 페이징 결과와 비교되는 정답은 1000 genomes에서 제공하는 gold-standard variant call[8]을 사용한다. 이는 하플로타입 페이징 알고리즘의 성능을 평가할 때 가장 많이 사용되는 신뢰도가 높은 데이터이다.

3.1 ProbHap의 품질정보 사용유무에 따른 정확도

ProbHap은 품질정보를 이용한 확률적모델을 사용한 하플로타입 페이징 알고리즘이다. 본 실험에서는 품질정보의 사용 유무에 따른 ProbHap의 정확도를 분석하여 품질정보의 필요성을 제시한다. 여기서는 품질정보의 개입을 하지 않도록 SNP matrix내의 품질 정보를 모두 동일한 값으로 설정하여 실험을 진행한다.

품질정보 사용 유무에 따른 ProbHap의 정확도를 (표 1)의 에서 나타내고 있다. 품질 정보를 사용한 경우가 사용하지 않은 경우보다 높은 정확도를 보여주고 있다. 이는 ProbHap에서 사용되는 확률적모델이 품질정보에 최적화되었다고 판단할 수 있다.

일반적으로 SNP matrix의 read들 간 서로 위치를 공유하지 공유하지 않고 끊어지는 부분이 존재한다. 대부분의 하플로타입 페이징 알고리즘은 이런 끊어지는 부분들을 기준으로 블록을 나누어 블록 별로 하플로타입 페이징을 진행한다. (표 1)은 품질정보를 사용한 경우와 사용하지 않은 경우 서로 더 높은 정확도를 가진 블록의 개수를 보여주고 있다. 품질정보를 사용한 경우가 사용하지 않은 경우보다 정확도가 높은 블록의 개수가 9배 이상인 것을 볼 수 있다.

SNP matrix상의 품질정보가 명확하게 구분되지 않았지만, 결과적으로는 품질정보를 사용하는 것이 하플로타입의 높은 정확도를 위한 타당성 있는 시도라 판단할 할 수 있다.

	정확도	정확도가 높은 블록 수
품질정보 적용	94.33%	128
품질정보 미적용	85.17%	13

(표 1) ProbHap의 품질정보 사용유무에 따른 결과

3.2 BOAhap의 품질정보 사용 유무에 따른 정확도

BOAhap은 베이지안 네트워크를 사용한 유전알고리즘을 이용하여 하플로타입 페이징을 하는 알고리즘이다. 유전알고리즘에서는 매 세대가 지날 때 마다 해들을 평가하기 위해 적합도 함수를 사용한다. BOAhap에서 사용하는 적합도 함수(그림 5)는 MEC 모델 기반의 계산방법을 이용하며, 이 과정에서 품질정보가 사용된다. 적합도 함수는 하플로타입 페이징 과정에서 만들어진 하플로타입을 SNP matrix의 각 read들 과 비교하여 일치하는 정도를 계산하여 적합도를 측정한다.

$$\text{적합도함수}(h) = \sum_i \min\left((D(h, M_i), D(h^*, M_i)\right)$$

$$D(h, M_i) = \sum_j (\delta(h_j, M_{ij})(1 - S_{ij}) + (1 - \delta(h_j, M_{ij}))S_{ij})$$

$$\delta(h_j, M_{ij}) = \begin{cases} 0 & h_j = M_{ij} \ or \ M_{ij} = '-' \\ 1 & h_j \neq M_{ij} \end{cases}$$

M_{ij} = i번째 $read$의 j번째 요소의 값
Q_{ij} = i번째 $read$의 j번째 요소의 품질
h = 하플로타입

(그림 5) 품질정보를 사용한 적합도 함수

본 실험에서는 BOAhap의 품질정보 사용유무에 따른 결과를 비교하여 품질정보가 결과의 정확도에 미치는 영향을 알아본다. ProbHap의 실험과 마찬가지로 SNP matrix내의 품질 정보를 모두 동일한 값으로 설정하는 방법으로 품질정보의 개입을 제한한다. 또한 BOAhap의 경우 베이지안 네트워크 사용으로 인해 메모리가 제한되어 SNP matrix 내에 가장 긴 블록 하나를 기준으로 실험을 진행하였다.

(표 2)에서 품질정보 사용 유무에 따른 BOAhap의 정확도를 보여주고 있다. 품질정보를 사용한 경우가 사용하지 않은 경우보다 약 2% 높은 정확도를 나타내고 있다. ProbHap과 비교하여 품질정보 사용유무에 따른 정확도 차이는 크지 않지만, 적합도 함수에서의 품질정보 사용이 정확도 향상에 도움이 된다는 것을 알 수 있다.

	정확도
품질정보 적용	83.75%
품질정보 미적용	81.47%

(표 2) BOAhap의 품질정보 사용유무에 따른 정확도

4. 결론

본 논문에서는 하플로타입 페이징에 사용되는 품질 정보의 문제점을 제시하고 품질 정보의 필요성을 파악하기 위해 이와 관련된 실험을 진행하였다. 결과는 품질정보를 사용하는 것이 사용하지 않는 것보다 더 좋은 결과를 보여주었다. 즉 하플로타입 페이징 알고리즘을 설계하는데 있어 품질 정보를 사용하는 것이 적절하다는 결론을 내릴 수 있다.

현재까지 베이지안 네트워크, HMM 모델, MEC 모델, 휴리스틱 등과 같은 다양한 접근방법을 이용한 하플로타입 페이징 알고리즘들이 제안되었다. 이들 중 품질정보를 사용하지 않는 알고리즘 중에 성능이 우수한 알고리즘들 또한 존재한다. 추후 이와 같이 품질정보를 사용하지 않고 좋은 정확도를 보여주는 알고리즘을 품질정보를 적용할 수 있도록 개선시키는 시도도 필요할 것으로 판단된다.

참고문헌

[1] 이홍란, 이제근, 장병탁, 황규백, 신수용 (2013), “하플로타입 페이징 알고리즘에 관한 조사연구 (A Survey on Computational Methods for Haplotype Phasing)”, 정보과학회논문지 : 소프트웨어 및 응용 제 40 권 제 11 호

[2] Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song, “Genotype and SNP calling from next-generation sequencing data”, Nat Rev Genet. 2011 June ; 12(6): 443 – 451. doi:10.1038/nrg2986.

[3] Yu Zhang, “A dynamic Bayesian Markov model for phasing and characterizing haplotypes in next-generation sequencing“, BIOINFORMATICS Vol. 29 no. 7 2013, pages 878 – 885 doi:10.1093/bioinformatics/btt065

[4] Volodymyr Kuleshov, “Probabilistic single-individual haplotyping”, BIOINFORMATICS Vol. 30 ECCB 2014, pages i379 – i385 doi:10.1093/bioinformatics/btu484

[5] ,이종찬, 나중채, “BOA를 이용한 하플로타입 추정”, 한국정보처리학회 2015년 추계학술발표대회 (2015-10)

[6] Sharon R. Browning, Brian L. Browning, “Haplotype phasing: Existing methods and new developments”, Nat Rev Genet. ; 12(10): 703 – 714. doi:10.1038/nrg3054.

[7]ftp:// ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878 NA12878/NIST_NA12878_HG001_HiSeq_300x

[8] 1000 Genomes FTP site, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/release/2010_07/trio/snps/

[9] X.-S. Zhang, L.-Y. Wu, and L. Chen, "Models and Algorithms for Haplotype Problem," Current Bioinformatics, vol.1, pp.105-114, 2006.