·MarSel : Large-scale Dataset에 대한 LD기반의 Marker 선택 시스템

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

MarSel : The LD-based Marker Selection System for the Large-scale Datasets

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

요 약

인간(human)에게 나타나는 다양성(variation)은 인체의 유전체(genome) 안에서 발생된 SNP(Single Nucleotide Polymorphism)에 의해 나타난다고 알려져 있다. 유전체내의 SNP과 다양성에 대한 연관 연구(Associate study)를 할 때에 약 30여억 개로 추정되는 염기서열(DNA sequence)을 모두 분석한다면 많은 비용과 시간을 필요로 할 것이다. 이런 비용과 시간을 줄이기 위해 적은 수의 대표 SNP(=tagSNP)을 찾는 연구가 현재 진행 중이다. 우리는 LD계수ID'I을 block 분할에 이용하여 생물학적인 의미를 부여한 후, 전산적인 최적해를 찾는 접근을 이용했다.

또한, 기존 연구에서는 large-scale data에 대한 처리가 불가능해서 chromosome의 일부분의 데이터에 대해서만 분석이 시도되었다. 더욱 광범위한 분석을 위해서 chromosome 단위의 처리가 필요하다. 우리는 chromosome단위의 SNP data를 한 번에 처리가 가능한 시스템인 MarSel를 구현하였다.

1. 서 론

인간(human)에게 나타나는 다양성(variation)은 인체의 유전자(genome) 안에서 발생된 SNP에 의해 나타난다고 알려져 있다. 유전자내의 SNP과 다양성에 대한 연관 연구(Associate study)를 할 때에 약 30여억개로 추정되는 모든 염기서열(DNA sequence)을 검사한다면 많은 비용과 시간이 소요된다. 이런 비용과 시간을 줄이기 위해 SNP들 중에 Marker로 사용되어지는 tagSNP의 수를 최소한으로 찾아내기 위한 많은 연구들이 진행 중이다. 그러나 기존 연구들 중에서는 단지 전산학적인(computational) 관점에서만 이 문제를 풀었다. 우리의 방법은 Linkage Disequilibrium(LD)을 고려하여 전산학적 접근법에 생물학적인 의미를 부여하였다.

또한 복잡한 질병(complex disease)의 경우 chromosome단위의 여러 SNP들이 조합되어 발생되어지는데 기존의 시스템은 large-scale data에 대한 처리가 불가능하여 chromosome내의 특정부분에서만 분석이 이루어졌다. 우리의 방법은 SNP수가 712,000개로 가장 많은 chromosome 1의 경우까지 한 번에 처리할 수 있다.

2. 알고리즘

2.1 MarSel의 구성

MarSel의 처리는 그림1처럼 총 4단계로 구분할 수 있다.

1) 1단계(입력) : 입력파일명과 threshold들을 입력하고, 입력되어지는 데이터의 sample수와 SNP수를 확인하는 단계.

2) 2단계(블록결정테이블 크기계산) : SNP수에 대하여 모든 경우의 블록크기를 측정해야 하는데 이때 ID'I계산을 통하여 가능한 block의 수를 제한하고, 결정된 블록 수의 크기로 블록결정테이블을 생성하는 단계.

3) 3단계(블록분할) : 모든 가능한 블록에 대하여 목적함수 f(·)를 계산을 하여 그 값을 동적프로그래밍 알고리즘을 통하여 optimal block들을 결정하는 단계

4) 4단계(tagSNP 선택) : optimal block들로부터 tagSNP들을 선택하는 단계.

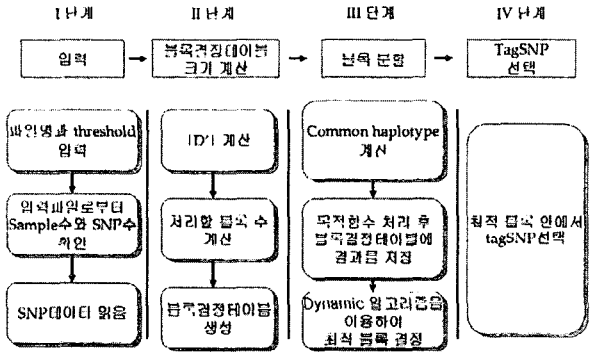


그림 1 MarSel 구성도

2.2 ID'I을 이용한 LD구간 선별

Linkage Disequilibrium(연관 불균형,LD)은 인접한 SNP간의 함께 유전된 경향을 나타내 주는 지표이다.

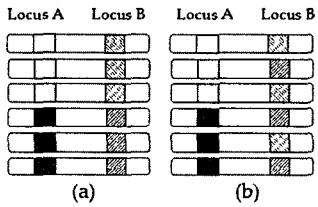


그림 2 LD 개념도

	A	a
B	n_{AB}	n_{aB}
b	n_{Ab}	n_{ab}
		N

그림 3 LD계산을 위한 분할표

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

그림 2에서 (a)는 locus A 와 locus B의 SNP은 함께 유전되는 부분이기에 LD가 있으며 recombination이 일어나지 않는다. (b)의 경우는 반대로 LD가 존재하지 않으며 recombination이 일어나는 부분인 것을 보여준다.

비교하려는 locus A, locus B의 2개의 SNP에서 Major, Miner를 locus A에서 A와 a, locus B에서 B와 b라고 정의를 내린다. 입력되는 모든 haplotype에 대하여 locus A와 locus B위치에서 AB, Ab, aB, ab인 경우의 frequency를 그림3처럼 계산하여 전체 sample수 N으로 나누어 P_{AB}, P_{Ab}, P_{aB}, P_{ab}를 구한다.

$$D = P_{AB} \times P_{ab} - P_{Ab} \times P_{aB}$$

$$|D|_{max} = \begin{cases} D>0 & \min(P_{aB}, P_{Ab}) \\ D<0 & \min(P_{AB}, P_{ab}) \end{cases}$$

$$|D'| = D / |D|_{max}$$

수식 1 |D'|의 정의

LD계수 |D'|은 위의 수식1처럼 정의를 하는데 D=0이거나 |D'|이 threshold로 정의한 α%미만일 때에는 생물학적으로 블록의 경계가 가능한 구간이라고 판단할 수 있다.

2.3 Entropy를 이용한 tagSNP selection

tagSNP은 블록 내에서 common haplotype들을 적은 수의 SNP으로 구분할 수 있는 SNP들을 의미한다. 우리의 방법에서는 tagSNP을 선택하기 위해 Entropy방법을 사용하였다. Entropy는 특정위치의 SNP(들)으로 common haplotype이 구분되어지는 척도를 나타낸다. 블록 내의 모든 SNP으로 구분되어지는 common haplotype에 대하여 standard entropy를 구한 후, 특정위치의 SNP(들)의 entropy를 계산하여 standard entropy에 가장 유사한 SNP을 tagSNP으로 인정한다. 위의 방법을 반복하여 tagSNP의 수를 늘려 entropy가 standard entropy의 threshold α%이내에 근접하면 그 때의 tagSNP수가 해당블록의 tagSNP수로 인정된다. entropy는 수식2와 같이 정의한다.

블록내의 특정 위치의 SNP(들)으로 인해 나누어진 n개의 common haplotype에서 i번째 common haplotype의 frequency를 A_i라고 하였을 때

$$Entropy = -\sum_{i=1}^{n} A_i \times \log_2 A_i$$

수식 2 entropy 정의

2.4 처리할 block수의 계산

MarSel은 dynamic method를 사용하여 모든 block에 대하여 계산을 한다. 하지만 대량의 SNP을 처리하기 위해 가능한 block에 대한 것을 줄여 주어진 조건하에 최대한 적은 계산을 하도록 설계하였다.

그림4는 가능한 블록을 결정하는 알고리즘이다. SNP a와 a+1의 |D'|이 LD threshold보다 작을 때 a+1은 Block endpoint table에 저장이 된다. 모든 SNP에 대하여 LD계산이 끝나면 Blockendpoint table로 부터 계산되어 Startpoint와 Endpoint table에 block의 시작점과 끝점이 각각 저장이 된다. Endpoint[x]일 때, Startpoint[0]부터 Startpoint[x]까지 각각 가능한 block을 이룬다.

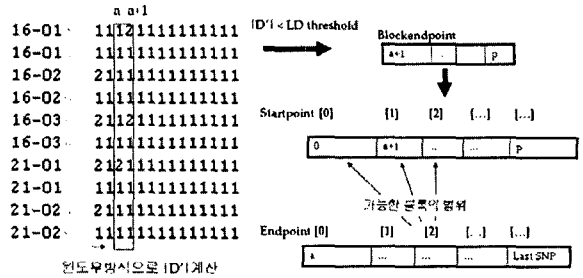


그림 4 가능한 블록결정 알고리즘

이렇게 가능한 블록을 계산해줌으로서 수식3처럼 계산량을 줄이는 효과를 갖게 설계되었다.

SNP : SNP수
A : LD구간의 수
B : LD구간에 속한 SNP수

$$Difference = \frac{(B-A)(2 \times SNP - (B-A)+1)}{2}$$

수식 3 가능한 블록 결정으로 생기는 계산량 차이

이렇게 결정된 가능한 block을 대상으로 common haplotype을 계산을 거쳐서 목적함수 f(·)을 계산하여 Decision table에 저장되어진다. 이 때 사용하는 목적함수는 수식4와 같다.

single common haplotype수 < (1-haplotype threshold)*sample number 일 때

$$f(\cdot) = \frac{\text{블록길이}}{common\,haplotype\text{수}}$$

수식4 MarSel의 목적함수

모든 가능한 블록에서 tagSNP을 구한다는 것은 계산량이 많고, 중복되는 값들이 많이 존재하여 common haplotype수를 구하는 것보다 비효율적이다. $\log_2$common haplotype수≤tagSNP수이기에 목적함수에서 tagSNP수를 common haplotype수로 대체하여 계산량을 줄였다.

3. 실험 환경 및 실험 data

3.1 실험 환경

MarSel v1.0의 성능을 비교하기 위해서 dynamic method를 사용한 HapBlock v3.0[3]과 greedy method를 사용한 HaploBlockFinder v0.7[4]을 이용하였다.

프로그램 명	시스템사양
MarSel v1.0	P4 3.2GHz(HT) 512MB WinXP
HapBlock v3.0	
HapBlockFinder v0.7	Xeon 550MHz(Dual) 768MB Linux

표 2 실험환경

프로그램별 시스템 사양은 표2에서 보여준다.

3.2 실험 data

MarSel의 성능 평가를 위해서 Patil[1]이 20명의 chromo some 21에서 찾은 24,047SNPs인 haplotype data를 사용하였다. 이 data를 이용하여 100,000SNPs, 120,000SNPs, 700,000SNPs의 haplotype data를 인공적으로 만들었다.

4. 결 과

4.1. 전산적인 접근법과 생물학적 접근법

비교대상	#Block	#tagSNP	Note
Kui Zhang(2002)[2]	2,575	3,582	전산적인접근법 Coverage 80% 적용
MarSel v1.0	1,840	3,851	생물학적인 접근법 LD threshold 0.8적용

*20 samples, 24,047SNPs Patil data 사용

표 3 전산적인 접근법과 생물학적 접근법의 결과차이

Kui Zhang(2002)에서 전산학적 접근법으로 Patil data를 이용하여 최소의 tagSNP을 갖는 block으로 나누었다. MarSel v1.0의 경우에 생물학적 접근법을 이용하여 전산적인 접근법보다 많은 tagSNP을 찾았지만, 적은 Block으로 나누어진 결과를 표에서 볼 수 있다. 이는 MarSel에서 찾은 tagSNP이 kui zhang의 방법보다 marker역할을 더 효율적으로 할 수 있다는 결과이다.

4.2. 생물학적 접근법을 이용한 프로그램별 Data처리량

large-scale haplotype을 처리 할 수 있다는 것은 인간의 다양성을 더욱 광범위하게 분석할 수 있다는 것이다. MarSel의 data 처리량을 측정하기 위해 100,000SNPs, 120,000 SNPs, 700,000SNPs인 인공데이터를 이용했다.

22개의 상동 염색체 중에 SNP이 가장 적은 것은 chromosome 21로서 121,567개이고, 가장 많은 것은 chromosome 1로서 712,040개이다. 이들의 chromosome 단위로 처리가 가능한지 100,000개, 120,000개와 700,000개의 haplotype data에 대하여 수행해보았다.

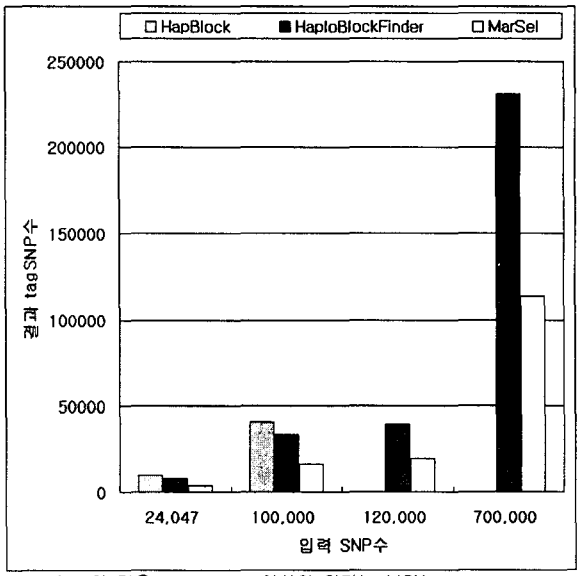
#SNP	MarSel v1.0		HapBlock v3.0		HaploBlockFinder v0.7	
	#Block	#tagSNP	#Block	#tagSNP	#Block	#tagSNP
24,047	1,840	3,851	4,135	9,754	3,453	7,826
100,000	7,761	16,228	17,378	40,472	14,303	33,138
120,000	9,226	19,305	수행불능		17,384	39,470
700,000	54,329	113,591	수행불능		101,976	231,238

*Sample수=20 LD threshold=0.8 coverage=0.8 tagSNP threshold=0.9

**HapBlock은 LD threshold대신에 fraction of Strong LD pair를 1로 입력

표 4 프로그램별 데이터처리량 비교

HapBlock v3.0의 경우에 120,000SNPs이상에 대해서는 처리하지 못하였고, MarSel v1.0과 HaploBlockFinder v0.7은 700,000SNPs의 large-scale haplotype data에 대해서 처리하였다. 같은 조건하에서 가장 긴 block(=가장 적은 수의 블록)과 적은 수의 tagSNP을 찾는 것이 좋은 결과이다. HaploBlockFinder의 경우에는 greedy algorithm 사용으로 근사해를 구한 결과이지만, MarSel의 경우 dynamic programming algorithm사용으로 최적해를 구했다는 가치가 있다. 그림5에서 HaploBlockFinder의 결과보다 50%미만의 tagSNP을 선택한 결과를 그래프 상으로 보여주고 있다.



*HapBlock의 경우 120,000SNPs이상의 처리는 불가능

그림 5 프로그램별 tagSNP selection결과

5. 결론 및 향후 연구과제

MarSel v1.0을 HapBlock v3.0, HaploBlockFinder v0.7과 비교한 결과 전산학적으로 HapBlock이 가장 optimal한 결과를 보였지만 생물학적인 의미를 부여했을 때 MarSel이 더 최적화된 결과를 보였다. 700,000SNPs을 처리한 MarSel v1.0을 통해 하나의 chromosome을 부분적으로 수행했던 association study가 chromosome 단위로 연구가 가능해짐으로서 더욱 광범위한 분석이 가능해지리라 기대한다.

앞으로 더 많은 유전정보를 다루기 위해 genotype 데이터가 필요하다. 이를 위해 reconstruction문제를 해결하기 위해 연구하여 genotype 데이터를 이용하여 tagSNP을 찾아내도록 MarSel을 발전시키겠다.

6. 참고문헌

[1]N. Patil et al., "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21," Science, 294:1719-23, 2001.

[2]Kui Zhang, Minghua Deng, Ting Chen, Michael S. Waterman, and Fengzhu Sun, A Dynamic Programming Algorithm For Haplotype Block Partitioning, PNAS, 2002(99): 7335-7339

[3]http://hto-b.usc.edu/~msms/HapBlock

[4]http://cgi.uc.edu/cgi-bin/kzhang/haploBlockFinder.cgi