LD-based tagSNP Selection System for Large-scale Haplotype and Genotype Datasets

대용량의 Haplotype과 Genotype데이터에 대한 LD기반의 tagSNP 선택 시스템

Sang Jun Kim[1], Sang Soo Yeo[1], Sung Kwon Kim[1*]

[1] School of Computer Science & Engineering, Chung -Ang University, Korea

*To whom correspondence should be addressed. E-mail: skkim@cau.ac.kr

Abstract

In the disease association study, the tagSNP selection problem is important at the view of time and c ost. We developed the new tagSNP selection system that ha s also facilities for the haplotype reconstruction and missing data processing. In our system, we improved biological meanings using LD coefficients as well as dynamic programming method. And our sys tem has capability of processing large -scale dataset, such as the total SNPs on a chromosome. We have tested our system with various dataset from daly et al., patil et al., HapMap Project, artificial dataset, and so on.

Introduction

인간(human)에게 나타나는 다양성(variation)은 인체의 유전자(genome)안에서 발생된 SNP에 의해 나타난다고 알려져 있다[1]. 유전체내의 SNP과 다양성에 대한 연관연구를 할 때에 약 30여억개로 추정되는 DNA sequence를 모두 분석한다면 많은 비용을 필요로 하게 된다. 이런 비용을 줄이기 위해 대표SNP (tagSNP)을 선택하는 문제에 대해서 연구가 되어오고 있다.

기존의 Kui Zhang의 HapBlock[2]과 같은 tagSNP Selection시스템들은 전산학적인 방법만으로 tagSNP을 selection 해왔다.

본 논문에서는 기존의 전산학적 접근법에 LD(Linkage Disequilibrium)을 고려하여 생물학적인 의미를 부여하였다.

또한 복잡한 질병(complex disease)의 경우 chromosome단위의 여러 SNP들이 조합되어 발생되는데 기존의 시스템은 large-scale data에 대한 처리가 불가능하여 chromosome의 관심 있는 특정부문만을 입력 받아 처리했었다.

본 논문에서 제안하는 시스템은 70만여개의 SNP에 대해서도 처리가 가능하고 genotype 데이터에 대한 처리도 Haplotype

본 연구는 한국 과학재단 목적기초연구 (R01-2003-000-11573-0)지원으로 수행되었음.

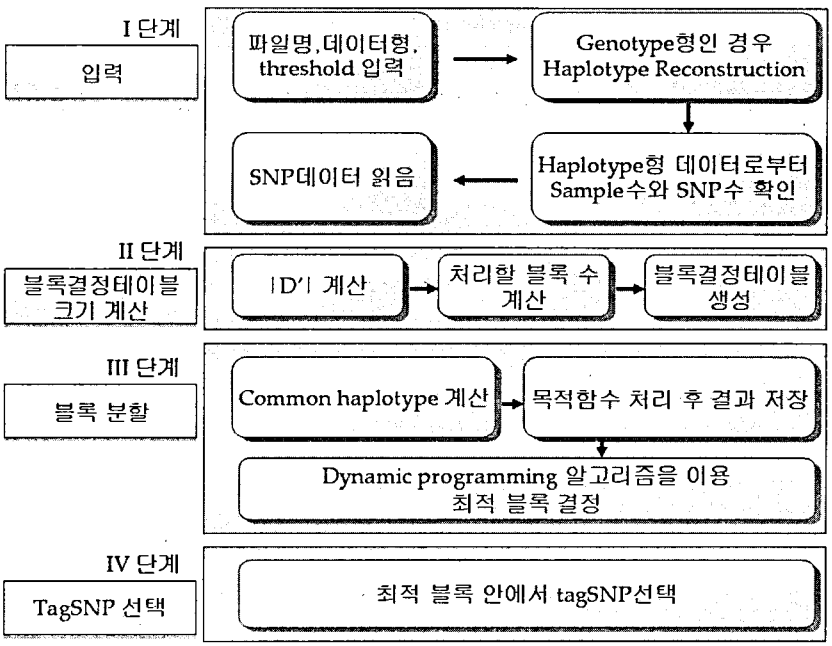

그림 0 시스템구성도

reconstruction을 하여 가능하기 때문에 chromosome단위의 광범위한 분석이 가능하다.

Methods

시스템구성

제안하는 시스템은 그림1에서 보듯이 총 4단계의 과정을 거쳐서 tagSNP을 선택하게 된다.

1단계: 입력 관련 처리

data형태와 입력파일명과 threshold들을 입력하고, 입력되는 data의 sample수와 SNP수를 확인하는 단계이다. data형태가 genotype data인 경우에는 phylogeny 방법[3]을 응용한 haplotype reconstruction단계를 거쳐 haplotype으로 추정 후에 2단계로 넘어간다.

2단계: 블록결정테이블 크기계산

SNP수에 대하여 모든 경우의 블록 크기를 측정해야 하는데 이때 |D'| [4]계산을 통하여 potential block의 수를 제한하고, 결정된 potential block수의 크기로 블록 결정 테이블을 생성하는 단계이다.

특히 |D'|계산을 통해 LD 블록을 나누어 입력 data에 포함된 missing data에 대하여 추측되는 값으로 대치시키는 과정이 포함된다.

3단계: 블록분할

모든 potential 블록에 대하여 목적함수 f(•)를 계산을 하여 그 값을 dynamic programming 알고리즘을 통하여 optimal block들을 결정하는 단계이다.

4단계: tagSNP 선택

3단계에서 결정된 optimal block들로부터 entropy 방법을 이용하여 tagSNPs를 선택 하는 단계이다.

|D'|을 이용한 LD block선별

Linkage Disequilibrium(연관 불균형, LD)은 인접한 SNP간의 함께 유전된 경향을 나타

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

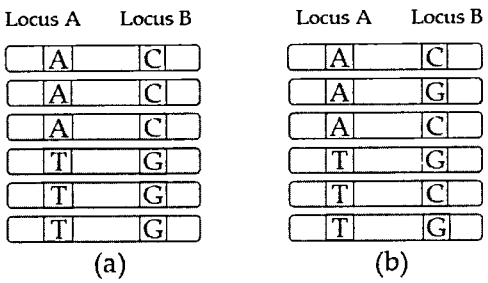

그림 1 LD개념도

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.

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

그림 2 대립유전자 빈도표

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

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

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

수식 1 |D'|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.

Potential 블록 결정

우리가 구현한 방법은 Dynamic program method를 사용하여 모든 block에 대하여 계산을 한다. 하지만 대량의 SNP을 처리하기 위해 가능한 block에 대한 것을 줄여 주어진 조건하에 최대한 적은 계산을 하도록 설계하였다. 이 때 적용한 것이 LD블록이다.

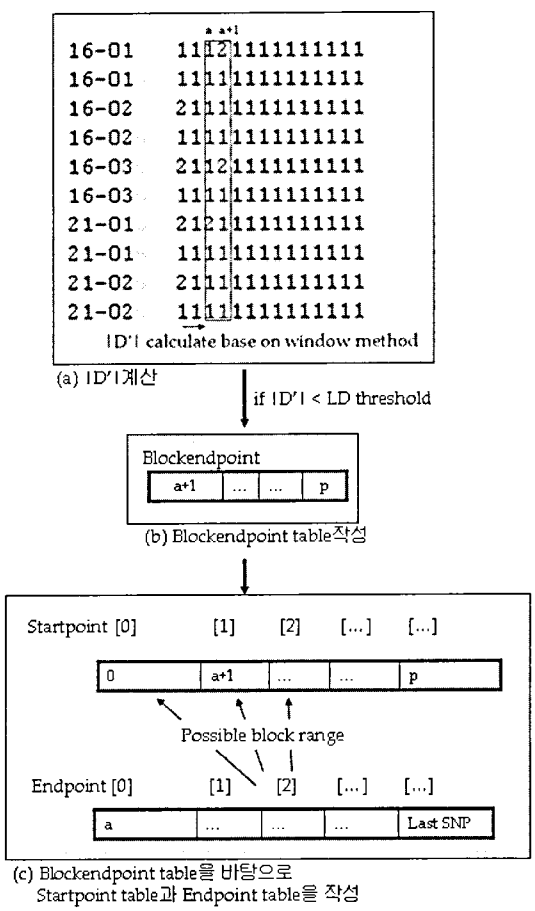

그림 3 Potential 블록결정 알고리즘

LD블록은 생물학적으로 recombination이 일어나는 부분으로 그 부분은 실제로 끊길 수 없다. 그래서 LD블록내의 블록들은 가능한 block들로 나누어 질 수 없고, 이 점을 이용하여 처리해야 할 블록의 수를 줄였다. 그

림4에서 potential 블록을 결정하는 알고리즘을 보이고 있다.

연속된 SNP a 와 SNP a+1의 |D'|이 LD threshold보다 작을 때 Blockendpoint table에 a가 저장된다. 모든 SNP에 대하여 LD계산이 끝나면 Blockendpoint table로부터 계산되어 Startpoint와 Endpoint table에 potential block의 시작점과 끝점이 각각 저장이 된다. Endpoint[x]일 때, Startpoint[0]부터 Startpoint [x]까지의 경우가 각각 potential 블록을 이룬다.

이렇게 결정된 potential 블록을 계산해줌으로써 수식2처럼 계산량을 줄이는 효과를 갖게 설계되었다.

SNP : SNP 수
A : LD 구간의 수
B : 모든 LD 구간에 속한 SNP수

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

수식 2 Potential 블록결정으로 생기는 계산량 차이

목적함수

위의 블록에서 목적함수를 계산하여 dynamic programming 알고리즘을 통하여 최소의 tagSNP을 갖는 블록으로 나누어서 전체 데이터의 tagSNP수를 계산하게 된다. 우리가 제시하는 목적함수는 common haplotype 수를 이용하는 방법인데 다음의 수식 3과 같다.

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

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

수식 3 목적함수정의

목적함수에서 조건은 common haplotype이 단독으로 존재하는 경우가 haplotype threshold로 정한 한계보다 많은 경우 해당 블록은 우리가 원하는 블록에서 제외시키기 위해 세웠다. 단독으로 존재하는 common haplotype이 많다는 것은 해당블록은 특정한 특징을 갖지 못하기 때문이다.

Entropy를 이용한 tagSNP selection

tagSNP은 블록 내에서 common haplotype의 구분을 적은 수의 SNP으로 정확하게 구분해줄 수 있는 SNP들을 의미한다.

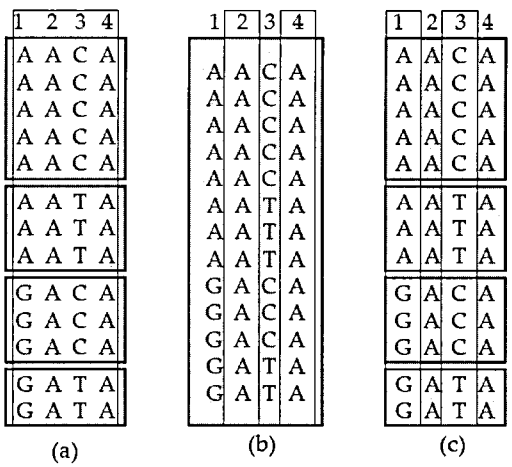

그림 4 tagSNP의 정의

그림 5에서 보듯이 (a)에서 4개의 SNP을 모두 비교해보면 총 4가지의 common haplotype으로 분류해볼 수 있다. 하지만 (c)에서처럼 2개의 SNP을 비교했을 때 4가지의 common haplotype을 구별해 낼 수 있다. (b)의 경우는 tagSNP을 잘못 선택했을 때 common haplotype을 잘못 구별한 예이다.
우리의 방법에서는 tagSNP을 정확하게 선택하기 위해 Entropy방법을 사용하였다.

Entropy는 특정위치의 SNP(들)으로 comm.-on haplotype이 구분되어지는 척도를 나타낸다. 그림 6의 (a)와 같이 특정 블록을 그림 5의 (a)처럼 블록 내의 모든 SNP으로 구분되는 common haplotype에 대하여 standard entropy를 구한 후, 그림 6의 (b)처럼 특정위치의 SNP(들)의 entropy를 윈도우형태로 계

산하여 standard entropy에 가장 유사한 SNP을 tagSNP으로 인정한다. tagSNP이 인정되었을 때의 entropy와 standard entropy의 tagSNP threshold α%이상 차이가 나면 인정된 tagSNP포함하여 위의 방법을 반복한다.

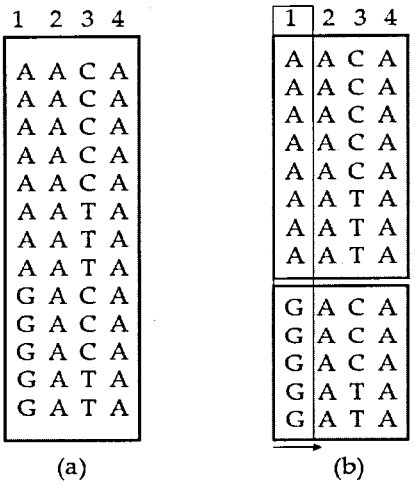

그림 5 entropy 계산 알고리즘

tagSNP의 수는 entropy가 standard entropy의 threshold α%이내에 근접할 때까지 증가되고 그 때의 tagSNP수가 해당 블록의 tagSNP수로 인정된다. Entropy는 수식 4와 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 4 entropy 정의

Missing Data의 대치법

우리가 사용하는 haplotype data와 genotype data는 여러 missing data들이 포함되어 있다. 우리는 missing data에 의해 나타나는 영향을 최대한 줄이기 위해 LD를 이용하였다. 우리가 제안한 방법에는 |D'|을 계산하는 단계가 있다. 이 때 missing SNP을 발견하면 앞,뒤의 SNP을 포함하여 3개의 SNP들로 |D'|을 계산한다. 계산식은 아래 수식 5와 같다.

계산한 수식을 바탕으로 가장 높은 연관성을 갖는 SNP으로 missing data를 대치해준다

$$|D'| = (P_{AB} \times |D'|_1) + (P_{BC} \times |D'|_2)$$

P_{AB} : 첫 번째,두 번째 SNP이 Major가 나왔을 경우의 확률
P_{BC} : 두 번째, 세 번째 SNP이 Major가 나왔을 경우의 확률
$|D'|_1$: 첫 번째, 두 번째 SNP간의 |D'|
$|D'|_2$: 두 번째, 세 번째 SNP간의 |D'|

수식 5 SNP 3개일 때 |D'|계산

Haplotype Reconstruction

PHASE Problem

Genotype data는 위상을 모르는 2개의 대립형질로 구성되어 있다. 위상을 모르기 때문에 2개의 haplotype data로 변환하기에 가능한 haplotype의 경우의 수가 그림 7에서 보듯이 $2^{SNP수}$ 가지이다.

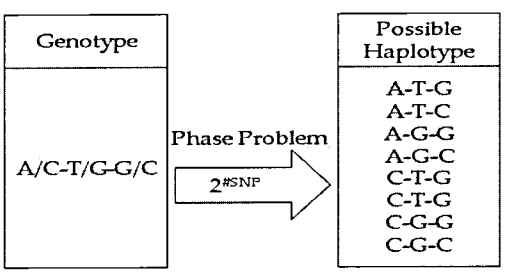

그림 6 Phase Problem

가능한 haplotype pair들 중에 실제 어떤 haplotype pair가 맞는 것인지를 추론하는 문제를 phase problem이라고 한다. 우리는 phase problem을 phylogeny 접근으로 해결하였다.

Perfect Phylogeny Haplotype

Perfect Phylogeny Haplotype은 가능한 haplotype pair를 기존의 통계학적 방법에만 의존하지 않고 계통발생론 입장으로 접근을 시도한 방법이다. 가능성 있는 haplotype들이 phylogeny tree에 일치되는 상황을 그림 8에

서 보인다.

Phylogeny tree에 일치를 시켜서 결정할 수 있는 경우의 수를 줄이고 maximum likelihood를 적용하여 haplotype pair를 결정한다.

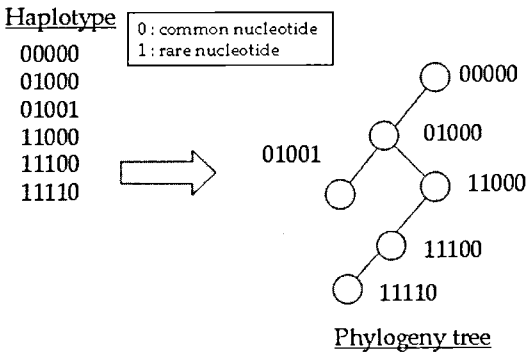

그림 7 Perfect Phylogeny Haplotype 개념도

Experiments and Results

실험환경

우리가 구현한 시스템(MarSel)의 성능평가를 위해 dynamic programming method를 이용한 HapBlock v3.0[5]과 greedy method를 이용한 HaploBlockFinder v0.7[6]을 사용하였다.

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

표 1 실험환경

실험 data

본 논문에서 제시하는 결과는 Patil data[7]와 100,000SNPs, 120,000SNPs, 700,000SNPs의 인공 data를 이용하여 실험하였다.

결과

표 2에서 Kui Zhang 의 전산학적인 방법과 비교했을 때 tagSNP수는 많았지만 block수가 적어 우리 방법으로 찾은 tagSNP이 전산학적 방법보다 더 긴 block으로 나누었음을 알 수 있다. 이것은 의미 있는 블록을 찾았다는 것에 의의가 있다.

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

*20 samples, 24,047SNPs Patil data 사용

표 2 Zhang과 MarSel v1.0의 결과비교

#SNP	MarSel v1.0		HapBlock v3.0		HaploBlock Finder 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 3 비교 프로그램간 data처리량 비교

표 3과 그림 2에서는 MarSel, HapBlock과 HaploBlockFinder의 데이터처리량에 대한 실험결과를 표와 그림으로 나타내고 있다. HapBlock의 경우 100,000SNPs에 이상에 대해서 처리가 불가능했으며, greedy method를 사용한 HaploBlockFinder가 찾은 tagSNP의 절반수준에 적은 tagSNP을 MarSel이 찾았다. 기존의 전산학적인 방법에 |D'|을 적용하여 생물학적인 접근을 시도한 우리의 방법이 데이터 처리량의 관점과 선택한 tagSNP수의 관점에서 가장 좋은 결과를 얻었음을 위의

결과에서 알 수 있다.

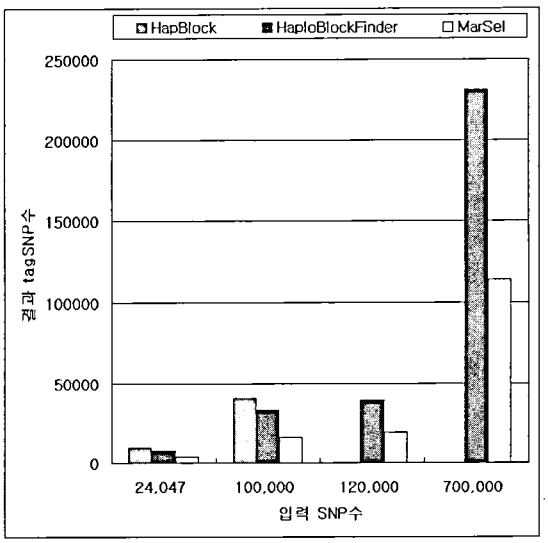

그림 8 비교 프로그램 별 data처리량 비교 결과

Discussion

우리는 genotype data와 haplotype data를 입력으로 받아서 optimal tagSNP selection을 수행하는 MarSel을 개발하였다. MarSel은 기존의 방법과는 달리 |D'|와 phylogeny method를 적용하여 생물학적 의미를 부여하였다.

또한 MarSel에 입력 가능한 SNP의 수 도 700,000개 이상으로써, chromosome단위의 연관 연구도 가능하게 되었다. 하지만 아직 여러 내부 모듈에 대한 최적화를 통해서 수행속도 향상이 필요하리라 본다.

Acknowledgements

본 연구는 한국 과학재단 목적기초연구 (R01-2003-000-11573-0)지원으로 수행되었음.

References

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

[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] Dan Gusfield. "Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions (extended abstract)" In *Proceedings of the 6th International Conference on Computational Molecular Biology (RECOMB2002)*, 2002

[4] R.C.Lewontin, "The interaction of selection and linkage. I. General considerations; heterotic models", *Genetics*, Vol.49, pp.49-67, 1964

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

[6] http://cgi.uc.edu/cgi-bin/kzhang/HaploBlock Finder.cgi

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