MarSel : 대용량 SNP 일배체형 데이터에 대한 연관불균형기반의 tagSNP 선택 시스템

김 상 준[†] · 여 상 수[††] · 김 성 권[†††]

요 약

최근 인간의 다양성과 SNP과의 연관연구에 드는 비용을 줄이기 위해서, 최소의 tagSNP을 선택하는 문제를 해결하기 위한 연구가 이루어지고 있다. 일반적으로 많은 수의 SNP들을 여러 블록으로 분할하여 각 블록 내에서 tagSNP을 선택하는 접근방법이 사용되고 있다.

본 논문에서 구현된 MarSel은 기존의 블록분할 접근 방법의 문제로 볼 수 있는 생물학적 의미의 부족을 해결하고자, 연관불균형(Linkage Disequilibrium, LD)의 개념을 도입한 시스템이다. 기존의 접근방법에서는 생물학적으로 재조합(recombination)이 일어나지 않는 연속된 구간에서도 여러 블록으로 나누어지는 문제가 생겼던 반면, MarSel에서는 연관불균형 계수 |D'|에 의해서 연속된 구간이 하나의 블록으로 유지된 상태에서 tagSNP을 선택하게 된다.

또한 MarSel에서는 각 블록 내에서 tagSNP을 선택 할 때에 엔트로피(entropy) 기반의 최적해 알고리즘을 이용함으로써 최소한의 tagSNP선택을 보장하게 되며, 기존의 구현된 시스템들보다 더 많은 양의 데이터를 효율적으로 처리할 수 있도록 구현되었기 때문에 염색체 레벨의 연관연구도 가능하게 해준다.

키워드 : 단일염기변이, 동적계획법, 일배체형, tagSNP선택, 연관불균형

MarSel : LD based tagSNP Selection System for Large-scale SNP Haplotype Dataset

Sang Jun Kim[†] · Sang-Soo Yeo[††] · Sung Kwon Kim[†††]

ABSTRACT

Recently the tagSNP selection problem has been researched for reducing the cost of association studies between human's diversities and SNPs. General approach for this problem is that all of SNPs are separated into appropriate blocks and then tagSNPs are chosen in each block.

MarSel in this paper is the system that involved the concept of linkage disequilibrium for overcoming the problem that the existing block partitioning approaches have short of biological meanings. In most approaches, the contiguous regions, which recombinations have not been occurred in, may be separated into several blocks. Otherwise, in MarSel, the each contiguous region is kept into one block using LD coefficient |D'| and then tagSNP selection step is performed.

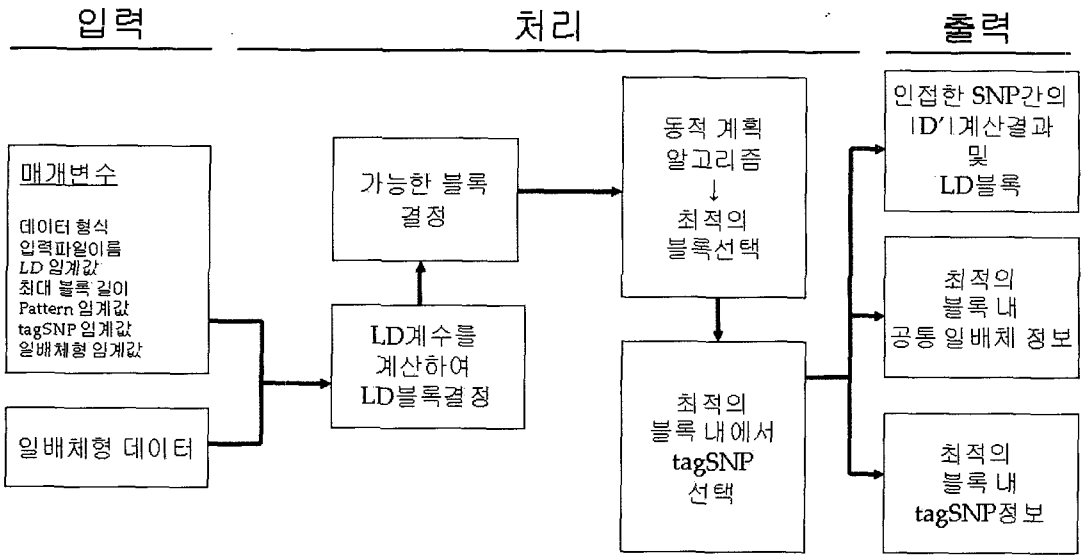
And MarSel guarantees the minimum tagSNP selection using entropy-based optimal selection algorithm when tagSNPs are chosen in each block, and enables chromosome-level association studies using efficient memory management technique when input is very large-scale dataset that is impossible to be processed in the existing systems.

Key Words : SNP, Dynamic Programming, Haplotype, tagSNP Selection, Linkage Disequilibrium

1. 서 론

현재 인간 유전체 프로젝트(human genome project)이후 밝혀진 염색체의 서열에 대한 기능을 연구하는 일배체형지도 프로젝트(hapmap project)가 수행되어지고 있다. 이를 통해 인종간 또는 개개인에게 나타나는 다양성, 특정질병유무, 특정약물에 대한 반응성 등이 유전체에 존재하는 변이(mutation)로 인해 나타난다고 밝혀지고 있다. 특히 이러한 변이 중에서 가장 흔하게 나타나는 SNP(단일염기변이, Single Nucleotide Polymorphism)은 인간 질병에 관여하는 것으로 약 1000bp의 염기당 1개의 빈도로 매우 빈번하게 관찰이 된다[1]. 일배체형지도 프로젝트의 일환으로 질병과 SNP간의 연관연구가 활발히 이루어지고 있는데 이때 약 30

※ 본 연구는 한국 과학 재단의 특정 기초 연구 과제로 지원받아 수행하였음 (R01-2003-000-11573-0).
† 준 회 원 : 중앙대학교 컴퓨터공학부 석사과정
†† 준 회 원 : 중앙대학교 컴퓨터공학부 박사과정
††† 종신회원 : 중앙대학교 컴퓨터공학부 교수
논문접수 : 2005년 6월 24일, 심사완료 : 2006년 1월 31일



(그림 1) MarSel의 구성도

여 억 개로 추정되는 염기 서열 에 존재하는 약 900여만 개의 SNP들을 모두 비교한다면 많은 비용과 시간이 소요된다. 이를 줄이기 위한 노력 중에 하나로 적은 대표SNP(tagSNP)을 찾는 문제에 대해서 많은 해결법들이 연구되고 있다. tagSNP은 전체 SNP중에 일배체형(haplotype)의 다양성을 잘 표현해주는 SNP(s)를 뜻한다. 기존의 연구에서는 계산적 접근법으로 최소의 tagSNP을 찾고 블록 분할하는 문제를 해결하였다. 하지만 Daly[6]의 주장에서 생물학적으로 재조합(recombination)이 일어나는 부분으로 블록을 이룰 수 있음을 보여주기 때문에 계산적 접근으로 해결한 tagSNP과 블록 분할은 실제 생물학적으로는 큰 의미가 없음을 알 수 있다. 우리의 방법은 연관불균형계수를 적용하여 계산적 접근법에 생물학적인 의미를 부여하였다.

또한 기존에 연구되어 구현된 대부분의 소프트웨어들은 700,000개 이상의 SNP 데이터를 처리할 수 있는 능력을 갖지 못하여 연관 연구하는데 있어 염색체의 특정부분을 분석하여 제한된 연구를 하였다. 700,000개는 가장 많은 SNP으로 알려진 염색체 1번의 SNP수로 이 이상의 SNPs를 처리할 수 있다는 것은 염색체 레벨 연관연구가 가능하다는 것을 보인다.

본 논문에서는 염색체 레벨의 연관 연구에 사용될 수 있는 MarSel을 구현하여 여러 실험을 통하여 성능평가를 하였으며, MarSel을 통해서 여러 질병과 SNP간의 연관연구가 더욱 효율적으로 이루어질 것으로 기대하고 있다.

2. MarSel의 구성

제안하는 시스템 MarSel은 그림1에 표현된 것과 같이 입력, 처리, 출력의 크게 3부분으로 나눌 수 있으며, 처리는 4부분으로 세분화 시킬 수 있다.

2.1 입력 단계

데이터 형식과 입력파일명과 LD 임계값, 최대블록길이, Pattern 임계값, tagSNP 임계값, 일배체형 임계값을 입력받고, 입력되는 데이터의 개체수와 SNP수를 확인하는 단계이다.

2.2 처리단계

우선 입력된 SNP에 대하여 모든 경우의 가능한 블록을 |D'|계산을 통하여 측정하고, 결정하게 된다. 결정된 가능한 블록수의 크기로 블록 결정 테이블을 생성한 후, 모든 가능한 블록에 대하여 목적함수 $f(\bullet)$를 계산을 한다. 이 값을 동적계획 알고리즘을 통하여 최적의 블록들을 결정한다. 이후 결정된 최적의 블록들로부터 엔트로피(entropy) 방법을 이용하여 tagSNPs를 선택하게 된다.

2.3 출력 단계

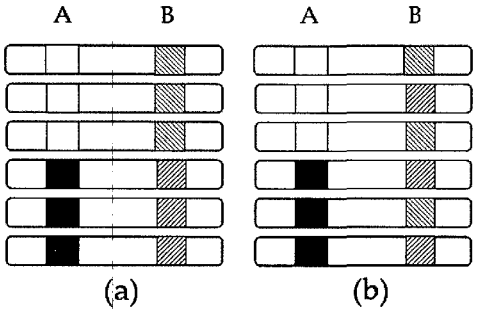
분석에 필요한 정보를 출력하는 부분으로 |D'|계산결과, 최적의 블록 내의 공통 일배체형(common haplotype)의 빈도수(frequency)정보, 최적의 블록 내의 tagSNPs정보를 출력한다.

3. 블록 분할

3.1 가능한 블록

가능한 블록은 생물학적으로 블록의 단위로서 적합한 블록을 의미한다. 계산적 접근법으로 최소의 tagSNP을 갖는 블록 분할을 한다면 생물학적으로 재조합이 일어나지 않는 부분이 블록 분할될 수 있다.이를 방지하기 위해서 우선 연관불균형 계수를 계산하고 인접한 SNP의 연관성을 고려하여 연관불균형 블록(LD 블록)을 결정하게 된다. 이렇게 결정된 연관불균형 블록의 조합들로 가능한 블록을 구성하여 공

통 일배체형의 수를 목적함수 f(•)를 통하여 제한하게 된다.



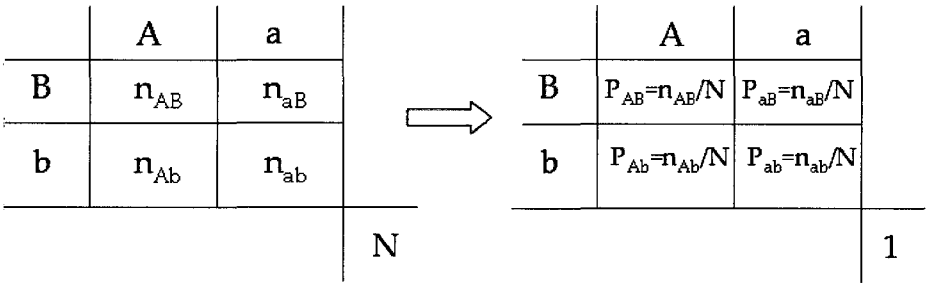
(그림 2) 연관불균형 개념도

3.1.1 |D'|를 이용한 연관불균형블록 결정

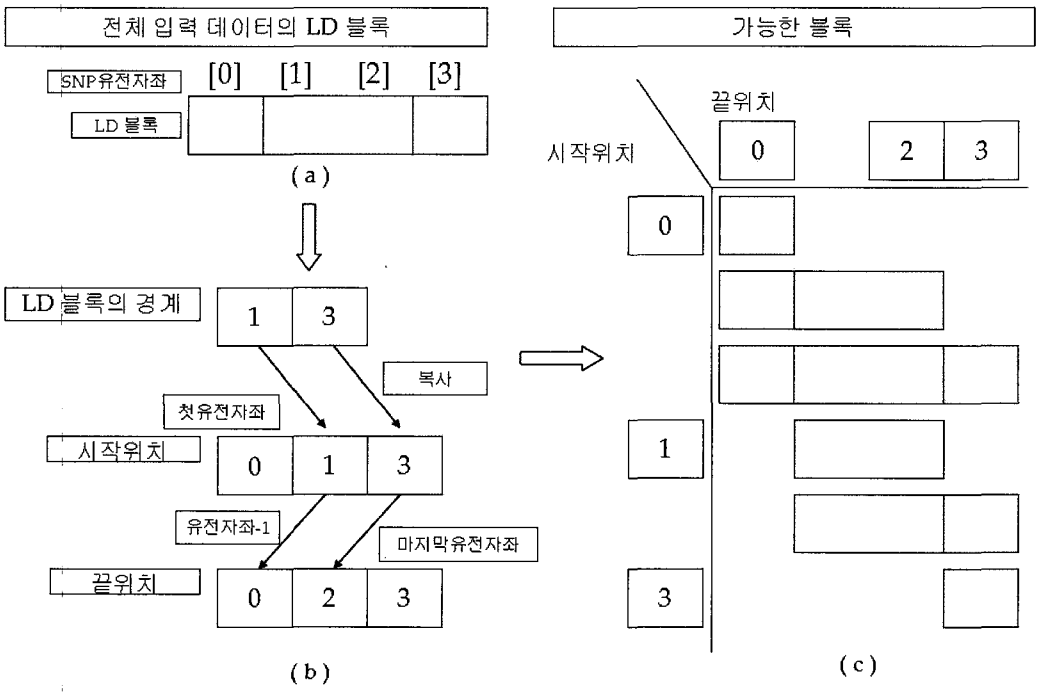
연관불균형(Linkage Disequilibrium, LD)은 인접한 SNP 간의 함께 유전된 경향을 나타내 주는 지표이다. (그림 2)에서 (a)는 A위치와 B위치의 SNP은 함께 유전되는 부분이기에 연관불균형이 있으며 재조합이 일어나지 않는다. (b)의 경우는 반대로 연관불균형이 존재하지 않으며 재조합이 일어나는 부분인 것을 보여준다.

비교하려는 A위치, B위치의 2개 SNP에서 주 대립 형질, 부 대립 형질을 A위치에서 A와 a, B위치에서 B와 b라고 각각 정의를 내린다. 입력되는 모든 일배체형에 대하여 A위치와 B위치에서 AB, Ab, aB, ab인 경우의 빈도수를 (그림 3)처럼 계산하여 전체 개체 수 N으로 나누어 P_{AB}, P_{Ab}, P_{aB}, P_{ab}를 구한다.

연관불균형계수 |D'|은 아래 수식 (1)~(3)처럼 정의된다 [2]. 만약 D=0이거나 |D'|이 연구자가 정하게 되는 LD 임계값(threshold) α보다 작은 경우 일 때에는 생물학적으로 재조합이 일어나는 위치로 간주하게 되고, 이곳을 연관불균형 블록의 경계 구간이라고 판단하게 된다. 여기서 LD 임계값 α는 연구자에 의해서 다양하게 설정이 될 수 있지만, 일반적으로 0.8 정도의 값이 의미 있는 LD 임계값으로 사용되고 있다[4, 5].



(그림 3) 대립유전자 빈도표



(그림 4) 가능한 블록 결정단계

$$D = P_{AB} \times P_{ab} - P_{Ab} \times P_{aB} \quad (1)$$

$$|D|_{max} = \begin{cases} \min(P_{aB}, P_{Ab}) & , \text{ if } D > 0 \\ \min(P_{AB}, P_{ab}) & , \text{ if } D < 0 \end{cases} \quad (2)$$

$$|D'| = \frac{D}{|D|_{max}} \quad (3)$$

3.1.2 연관불균형 블록의 경계를 이용한 가능한 블록결정

앞 절의 방법으로 전체 입력 데이터에 대해 연관불균형 블록이 나누어지면 연관불균형 블록들의 시작점과 끝점의 유전자좌 값이 시작 위치와 끝 위치에 각각 입력된다. 입력했던 최대블록길이의 범위 안에서 시작 위치와 끝 위치의 조합으로 가능한 블록이 결정되어진다.

(그림 4)는 4개의 SNP데이터에 대해서 가능한 블록 결정의 예를 보여준다. (그림 4)(a)는 전체 입력 데이터의 연관불균형 블록을 보인다. 2번째 블록이후의 시작점들은 LD 블록의 경계에 저장되어지는데 (그림 4)(b)에 1과 3의 유전자좌가 저장된 것을 보여준다. 이를 근거하여 시작 위치는 첫 유전자좌의 0이 저장된 뒤 연관불균형 블록의 경계에 저장된 값들이 저장되어진다. 그리고 끝 위치에는 연관불균형 블록의 경계에 저장된 값에서 -1이 된 값들이 저장되고 마지막 유전자좌가 저장된다. 이후에 (그림 4)(c)와 같이 조합을 이루어 가능한 블록을 결정하게 된다.

3.2 최적의 블록

최적의 블록이란 가능한 블록 중에서 1)최소의 tagSNP을 가질 것, 2)긴 블록일 것이라는 2가지 조건을 만족하여 분할된 블록들이다. 최소의 tagSNP을 갖기 위해 가능한 블록 중에서 공통 일배체형이 적게 나타나는 경우를 선택하게 된다. 공통 일배체형이 적게 나타난다는 것은 특정 블록 내에서 특정한 다양성의 구분을 비교 할 수 있다는 것이다.

3.2.1. 목적함수 f(•)

목적함수에 대해서 설명하기 전에 먼저 수식 (4)와 수식 (5)에 관련된 인자를 아래와 같이 정의한다.

t : 일배체형의 임계값
n : 샘플의 개수
b : 블록의 길이
h : 공통 일배체형의 개수
s : 단일 일배체형의 개수
m : tagSNP의 개수

우리 방법에서 가능한 블록 중에서 최적의 블록을 결정하기 위해서는 최소의 tagSNP수와 긴 블록이란 2가지 조건을 만족해야 한다. 이런 목적을 계산하기 위해 목적함수인 f(•)를 만들었다. f(•)에는 한 가지 제한사항이 있다. 특정 블록 내에서 한 개체에서만 나타나는 단일 일배체형의 빈도가 전체 개체수내에서 1-일배체형의 임계값%이내여야 한다는 것이다. 단일 일배체형은 가능한 블록에서 특정한 특징을 갖지 못하기 때문에 블록으로서 의미를 부여하지 않겠다는 의미를 담고 있다.

이런 제한을 갖고 시스템의 계산량을 줄이기 위한 목적을 갖고 만든 함수는 다음 수식 (4)와 같다.

$$f(\cdot) = \frac{b}{h} \quad , \text{ if } s < (1-t) * n \quad (4)$$

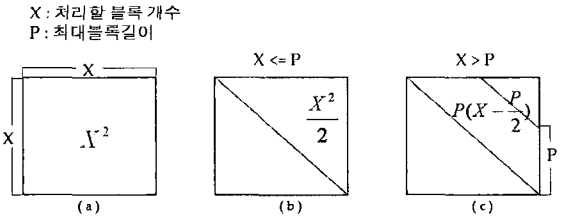
tagSNP수와 공통 일배체형수는 수식 (5)와 같은 관계를 갖기에 tagSNP수를 공통 일배체형수로 대치시킬 수 있었다.

$$\log_2 h \leq m \quad (5)$$

그래서 가능한 블록들을 대상으로 공통 일배체형 구분 함수를 이용하여 최적의 블록을 결정하고 결정된 최적의 블록 안에서 tagSNP 선택 함수를 사용하였다.

3.2.2 동적계획 알고리즘을 위한 배열할당

동적계획 알고리즘은 알고리즘 특성상 모든 가능한 경우에 대해서 처리한다. 그래서 (그림 5)(a)처럼 X개의 가능한 블록에 대해 처리하기 위해서는 X^2 크기의 메모리 공간을 차지한다. 하지만 실재 최대블록길이(P)의 제한으로 (그림 5)(b)와 (그림 5)(c)처럼 X^2의 절반 이하의 공간만 사용하게 된다. 그래서 메모리가 낭비되는 문제는 대용량 데이터를 처리하기 위해서는 해결해야 하는 문제이다. 우리는 인덱스(index)를 사용하여 X^2크기의 배열을 할당하는 대신에 최대블록길이(P) 에 따라 P*(X-P/2)크기 혹은 $X^2/2$ 크기의 배열을 할당하여 대용량 데이터처리문제를 해결하였다.



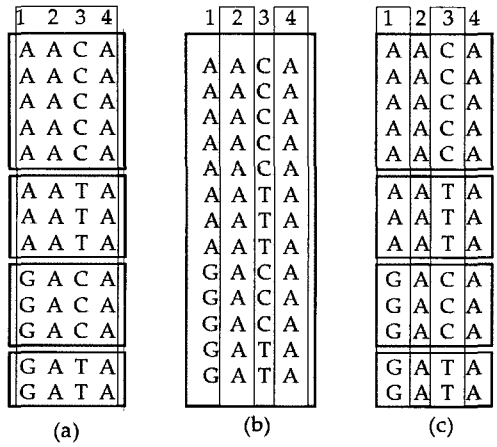
(그림 5) 메모리할당의 문제

4. tagSNP 선택

4.1 tagSNP

tagSNP은 전체 SNP중에 일배체형의 다양성을 표현해줄 수 있는 SNP들을 의미한다. (그림 6)(a)는 전체 SNP으로 일배체형의 다양성을 표현한 경우이고, (그림 6)(b)는 일배체형의 다양성을 표현하지 못한 SNP이 선택된 경우, (그림 6)(c)의 경우는 일배체형의 다양성을 제대로 표현한 SNP을 선택했을 경우를 나타내고 있다.

(그림 6)에서 보듯이 적절한 tagSNP의 선택은 일배체형의 다양성 구별에 드는 비용을 줄일 수 있기에 우리는 엔트로피를 이용하여 최소의 tagSNP 선택 문제를 해결하였다.



(그림 6) tagSNP 선택에 의한 일배체형의 다양성 표현

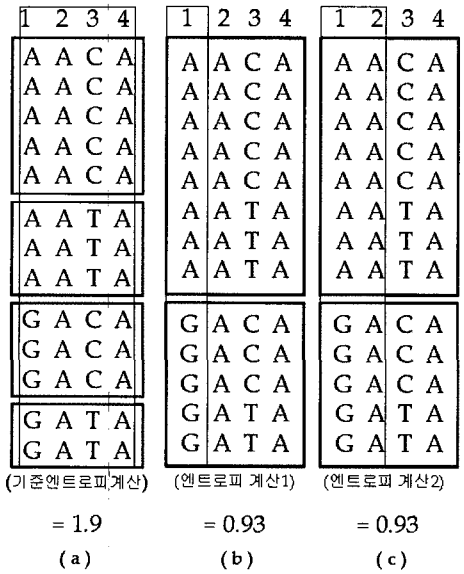
4.2 엔트로피를 이용한 tagSNP 선택

엔트로피는 특정위치의 SNP(들)으로 일배체형이 구분되어지는 척도로서 정의는 수식 (6)[3]과 같다.

$$Entropy = \sum_{i=1}^{n} (A_i \log_2 A_i), \quad (6)$$

A_i는 i번째 공통 일배체형의 빈도수

처리는 우선 (그림 7)(a)에서 보이듯 최적의 블록 안에 있는 모든 SNP에 대하여 비교하여 일배체형의 다양성별로 구분을 지어 기준 엔트로피를 계산한다. 그 후 (그림 7)(b)처럼 블록 안에서 윈도우 방식으로 특정한 1개의 SNP별로 일배체형의 다양성별로 구분을 짓고 엔트로피를 계산하는데 그 중에 기준 엔트로피와 가장 근접한 경우의 SNP을 선택한다. 선택된 SNP을 포함 시키고 또 다른 SNP들을 바꾸어 가며 (그림 7)(c)처럼 엔트로피를 계산하여 기준 엔트로피에 tagSNP 임계값 수준의 범위 안에 들 때까지 반복을 하여 tagSNP을 선택하게 된다.



(그림 7) 엔트로피 처리 단계

5. 실험 환경 및 실험 데이터

5.1 실험 환경

MarSel의 성능을 비교하기 위해서 동적계획법(dynamic programming)을 사용한 HapBlock v3.0[4]과 탐욕(greedy) 방법을 사용한 HaploBlockFinder v0.7[5]을 이용하였다. 프로그램별 시스템 사양은 <표 1>에서 보여준다.

〈표 1〉 실험환경

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

5.2 실험 데이터

MarSel의 성능 평가를 위해서 Daly 데이터[6], Patil 데이터[7], 인공 데이터를 사용하였다. Daly 데이터는 103명의 크론병(Crohn's disease) 환자와 그들의 부모 유전자형 데이터(genotype data)이고, Patil 데이터는 20명의 21번 염색체에서 찾은 24,047SNPs인 일배체형 데이터이다. 인공 데이터는 Patil 데이터를 이용하여 100,000SNPs, 120,000SNPs, 700,000SNPs의 일배체형 데이터를 만들었다.

6. 결 과

6.1. 전산적인 접근법과 생물학적 접근법

Kui Zhang(2002)은 전산학적 접근법으로 Patil 데이터를 이용하여 최소의 tagSNP을 갖는 블록으로 나누었다[8]. MarSel의 경우에는 연관불균형을 이용한 생물학적 접근법을 적용하여 전산적인 접근법보다 많은 tagSNP을 찾았다. 하지만 <표 2>와 <표 3>을 비교해 보면 MarSel이 Kui Zhang의 결과보다 더 적은 블록으로 분할한 것을 알 수 있다.

〈표 2〉 HapBlock(2002)의 tagSNP수 및 블록 분할의 결과

방법	블록단위 SNP 수	블록의 수	비율(%)	tagSNP 수	비고
HapBlock (2002)	>10	742	28.8	3,582	80% cover-age
	3-10	909	35.3		
	<3	924	35.9		
	Total	2,575	100		

〈표 3〉 MarSel의 tagSNP수 및 블록 분할의 결과

방법	블록단위 SNP 수	블록의 수	비율(%)	tagSNP 수	비고
MarSel	>10	845	45.9	3,921	80% coverage LD 임계값 0.8
	3-10	920	50.0		
	<3	75	4.1		
	Total	1,840	100		

블록단위 SNP수를 <표 2>와 <표 3>에서 비교해보면 HapBlock(2002)의 경우는 3bp미만인 경우와 3-10bp인 경우가 70%가 넘는 경우이지만 MarSel의 경우 3-10bp인 경우와 10bp이상인 경우가 95.9%로 긴 블록으로 분할 된 것을 볼 수 있다. 즉, MarSel은 재조합이 일어나지 않는 생물학적으로 의미 있는 블록들을 보호하며 나누었으며, HapBlock (2002)의 경우 재조합이 일어나지 않는 부분이 나누어져 생물학적 의미 있는 블록을 나누었음을 의미한다.

6.2 MarSel의 tagSNP의 표지자의 역할

(그림 8)은 Daly 데이터에서 부모의 데이터 중 15명의 환자인 부모의 유전자형 데이터와 15명의 정상인 부모의 유전자형 데이터를 일배체형 재구성(haplotype reconstruction)하여 총 60 개체를 갖고 tagSNP을 선택하고, 선택된 tagSNP을 갖고 같은 데이터의 일배체형의 다양성을 분별해보았다. 총 12 블록에서 4개의 블록에서는 100% 일치, 6개의 블록에서는 90%이상 일치, 2개의 블록에서 90%미만의 일치율을 보였다. 즉, 위의 실험은 MarSel이 선택한 tagSNP은 일배체형의 다양성을 구별할 수 있는 표지자(marker) 역할을 충분히 할 수 있다는 것을 증명하였다.

6.3 생물학적 접근법을 이용한 프로그램별 데이터 처리량

대규모 일배체형 데이터를 처리 할 수 있다는 것은 인간의 다양성을 더욱 광범위하게 분석할 수 있다는 것이다. MarSel의 데이터 처리량을 측정하기 위해 100,000SNPs, 120,000 SNPs, 700,000SNPs인 인공데이터를 이용했다.

22개의 상동 염색체 중에 SNP이 가장 적은 것은 21번 염색체로써 121,567개이고, 가장 많은 것은 1번 염색체로써 712,040개이다. 이들의 염색체 레벨로 처리가 가능한지 100,000개, 120,000개와 700,000개의 일배체형 데이터에 대하여 수행해보았다.

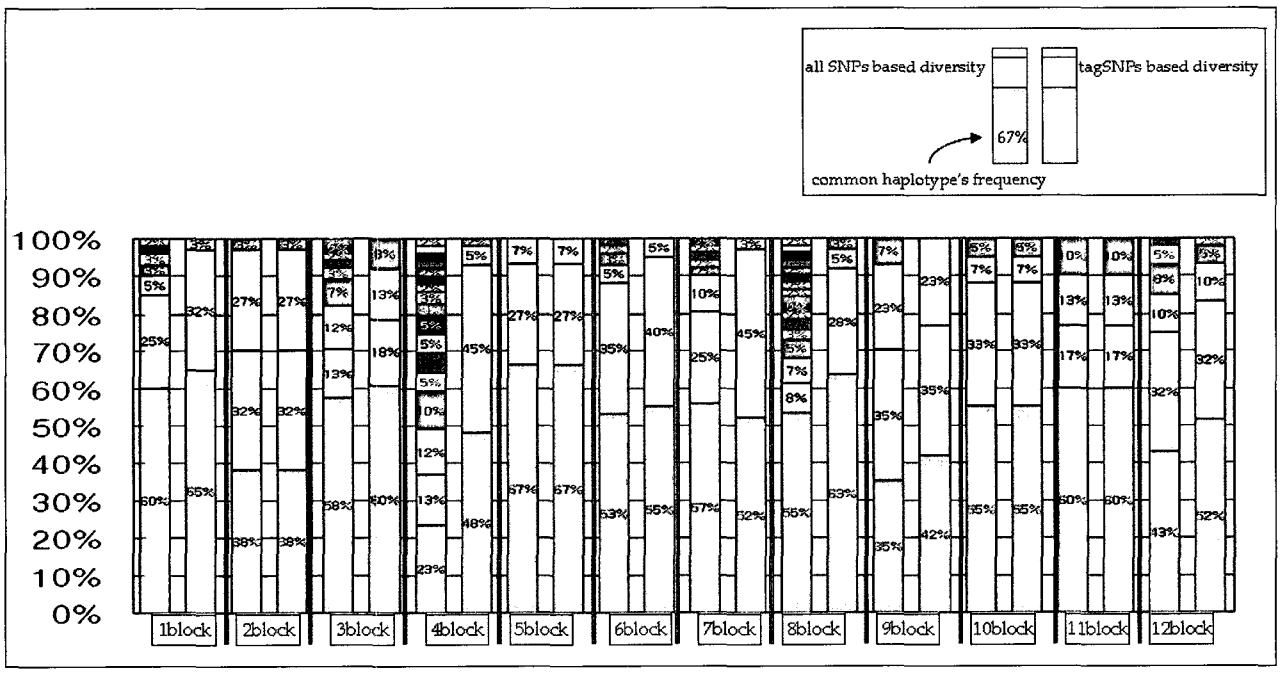
<표 4>에서는 HapBlock v3.0의 경우에 SNP 120,000개 이상에 대해서는 처리하지 못하였고, MarSel과 HaploBlockFinder v0.7은 SNP 700,000개의 대규모 일배체형 데이터에 대해서도 처리한 것을 볼 수 있다. 같은 조건하에서 가장 긴 블록(=가장 적은 수의 블록)과 적은 수의 tagSNP을 찾는 것이

〈표 4〉 프로그램별 데이터 처리량

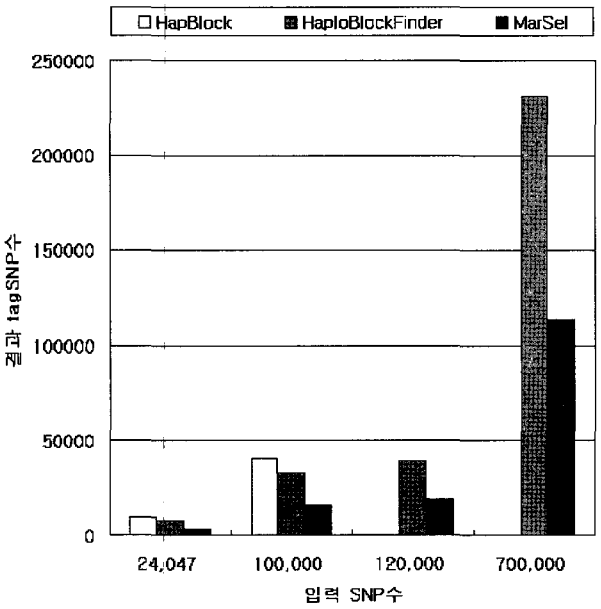
처리 SNP수	MarSel		HapBlock v3.0		HaploBlockFinder v0.7	
	블록 수	tagSNP 수	블록 수	tagSNP 수	블록 수	tagSNP 수
24,047	1,840	3,851	4,135	9,754	3,453	7,826
100,000	7,761	16,228	17,378	40,472	14,303	33,138
120,000	9,226	19,305	수행불능		17,384	39,470
700,000	54,329	113,591	수행불능		101,976	231,238

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

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



(그림 8) MarSel이 선택한 tagSNP의 분별성



*HapBlock의 경우 120,000SNPs이상의 처리는 불가능

(그림 9) 프로그램별 tagSNP selection결과

좋은 결과이다. HaploBlockFinder의 경우에는 탐욕 알고리즘을 사용하여 근사해를 구한 결과이지만, MarSel의 경우 동적계획 알고리즘의 사용으로 최적해를 구했다는 가치가 있다. (그림 9)에서 HaploBlockFinder의 결과보다 50%미만의 tagSNP을 선택한 결과를 그래프 상으로 보여주고 있다.

7. 결론 및 향후 연구 과제

본 연구를 통해서 인간의 다양성과 SNP간의 연관연구에 대한 비용을 줄이기 위해 최소의 tagSNP을 선택하는 시스템 MarSel을 개발하였다. MarSel은 기존의 방법에 비해서 일배체형 데이터를 생물학적으로 의미가 있는 블록으로 분할하였고, 블록 안에서 선택된 tagSNP의 수는 기존의 프로그램들의 결과보다 50%나 줄이면서도 일배체형의 다양성 표현 능력은 기존 프로그램들과 같은 수준을 보여준다. 또한 대용량의 데이터를 처리함으로서 기존의 염색체 일부의 연관연구에서 염색체 단위의 연관연구가 가능해졌다.

현재 MarSel은 일배체형으로 처리 완료된 데이터만을 처리할 수 있지만, 유전자형 데이터를 일배체형으로 재구성하는 연구와 이를 MarSel에 추가 구현하는 작업이 진행 중에 있다. 이것이 완료되면 인간 유전체 프로젝트로 생성된 많은 유전자형 데이터를 직접 MarSel의 입력 데이터로 넣어서 처리하는 것도 가능해진다.

참 고 문 헌

[1] J. I. Bell, "Single Nucleotide Polymorphisms and Disease Gene Mapping," Arthritis Research, Vol.4, pp.s273-s278, 2002.

[2] R. C. Lewontin, "The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models," Genetics, Vol.49, pp.49-67, 1964.

[3] R. Mott, "Marker Selection by Maximum Entropy," http://www.well.ox.ac.uk/~rmott/SNPS, Wellcome Trust Centre for Human Genetics, University of Oxford, 2003.

[4] K. Zhang, Z. Qin, T. Chen, J. S.. Liu, M. S. Waterman and F. Sun, "HapBlock: Haplotype Block Partitioning and Tag SNP Selection Software using a Set of Dynamic Programming Algorithms," Bioinformatics, Vol.21(1), pp.131-134, 2003.

[5] K. Zhang and L. Jin, "HaploBlockFinder: Haplotype Block Analyses," Bioinformatics, Vol.19, No.10, pp.1300-1301, 2003.

[6] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson and E. S. Lander, "High-Resolution Haplotype Structure in the Human Genome," Nature Genetics, Vol.29, No.2, pp.151-158, 2001.

[7] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. N. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. A. Fodor, D. R. Cox, "Blocks of Limited Haplotype Diversity Revealed by High-Resolution Scanning of Human Chromosome 21," Science, Vol.294, pp.1719-1723, 2001.

[8] K. Zhang, M. Deng, T. Chen, M. S. Waterman and F. Sun, "A Dynamic Programming Algorithm For Haplotype Block Partitioning," Proceedings of the National Academy of Sciences (PNAS), Vol.99, No.11, pp.7335-7339, 2002.

김 상 준

e-mail : jjuns@alg.cse.cau.ac.kr

2003년 중앙대학교 식품공학과 및 정보시스템학과(학사)

2005년 중앙대학교 대학원 컴퓨터공학과 (석사)

관심분야 : 알고리즘, 생물정보학

여 상 수

e-mail : ssyeo@alg.cse.cau.ac.kr

1997년 중앙대학교 컴퓨터공학과(학사)

1999년 중앙대학교 대학원 컴퓨터공학과 (석사)

2005년 중앙대학교 컴퓨터공학과(박사)

관심분야 : 알고리즘, 암호프로토콜, 생물정보학, RFID 보안

김 성 권

e-mail : skkim@cau.ac.kr

1981년 서울대학교 계산통계학과(학사)

1983년 한국과학기술원 전산학과(석사)

1983년~1985년 목포대학교 자연과학대학 전산통계학과 전임강사

1990년 미국 University of Washington Computer Science & Engineering (박사)

1991년~1996년 경성대학교 이과대학 전산통계학과 조교수

1996년~현재 중앙대학교 컴퓨터공학부 교수

관심분야 : 생물정보학, 암호응용 및 정보보호, 계산기하학 및 응용