http://dx.doi.org/10.3745/KTSDE.2013.2.10.697

Solving the Haplotype Assembly Problem for Human Using the Improved Branch and Bound Algorithm

Mun-Ho Choi† · Seung-Ho Kang†† · Hyeong-Seok Lim†††

ABSTRACT

The identification of haplotypes, which encode SNPs in a single chromosome, makes it possible to perform haplotype-based association tests with diseases. Minimum Error Correction model, one of models to computationally assemble a pair of haplotypes for a given organism from Single Nucleotide Polymorphism fragments, has been known to be NP-hard even for gapless cases. In the previous work, an improved branch and bound algorithm was suggested and showed that it is more efficient than naive branch and bound algorithm by performing experiments for *Apis mellifera* (honeybee) data set. In this paper, to show the extensibility of the algorithm to other organisms we apply the improved branch and bound algorithm to the human data set and confirm the efficiency of the algorithm.

Keywords : Haplotype Assembly Problem, Minimum Error Correction Model, Branch and Bound Algorithm

개선된 분기한정 알고리즘을 이용한 인간 유전체의 일배체형 조합문제 해결

최 문 호† · 강 승 호†† · 임 형 석†††

요 약

인간의 한쪽 염색체상에 나타나는 SNP의 서열인 일배체형을 식별해내면 효과적인 유전질병 연관검사를 할 수 있다. 주어진 SNP 단편들로부터 계산적인 방법으로 한 쌍의 일배체형을 조합하기 위해 제시된 모델 중 하나인 최소오류수정 모델은 단편에 손실이 없는 경우조차 NP-hard임이 증명되었다. 기존의 분기한정 알고리즘은 많은 계산시간을 요구함에 따라 실제 응용에 사용하기 어려웠다. 그러나 최근에 개선된 분기한정 알고리즘이 제시되었고, 꿀벌(*Apis mellifera*)의 유전자형 데이터를 대상으로 성능을 분석해봄으로써 개선된 알고리즘이 기존 분기한정 알고리즘보다 효율적임을 보였다. 본 논문에서는 인간의 유전자형 데이터를 대상으로 개선된 분기한정 알고리즘을 적용해 일배체형 조합문제를 수행한다. 실험을 통한 성능분석 결과, 개선된 분기한정 알고리즘이 인간 유전체에 대해서도 성공적으로 적용됨을 확인함으로써 다양한 생명체의 일배체형 조합문제에 적용 가능함을 보인다.

키워드 : 일배체형 조합문제, 최소오류수정모델, 분기한정 알고리즘

1. 서 론

대규모 염기서열 분석기술을 포함한 유전자 분석기술 발전은 인간 유전체 사업(Human Genome Project)의 완성이라는 결실을 낳았고, 질병 또는 약물반응 등에 관련된 유전자를 규명하기 위해 인간의 염색체에 존재하는 유전변이형에 대한 많은 분석과 기술개발이 이루어졌다. 또한 다량의 유전변이형 정보를 데이터베이스화 하려는 노력들이 International HapMap Project[1] 등을 통해 성과를 내고 있다. 이에 따라 각 개인의 유전정보 차이를 이용한 질병 예측, 진단 및 예방을 꿈꾸는 맞춤의학이 현재 유전체 연구 분야의 중심이 되고 있다.

인간 유전체에서 유전적 변이를 가장 풍부하게 보여주는 유전 마커로 단일염기다형성(SNP; Single Nucleotide Polymorphism)이 대표적이다. SNP의 차이는 개별 개체의 발현 형질의 차이와 직접적인 관련이 있는 것으로 밝혀졌으며, 특히 유전과 관련이 있는 질병 연구에 중요한 의미를 갖고 있다. 따라서 개인 간 SNP의 차이를 유형화하고 특정 질병과의 연관성을 분석하면 개인들의 질병 발생 가능성을 사전에 알 수 있다. 또한 특정 약물이나 치료 방법에 반응하는 개인의 양상도 각자 다르게 나타나는데, 이러한 약물이나 치료 방법과 SNP 유형의 관계를 규명하게 되면 맞춤의학의 실현이 가능하게 될 것이다. 이처럼 SNP는 생물학, 의학, 약학 등 여러 분야에서 중요한 의미를 갖는 유전 마커이다.

※ 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2010-0023234).
† 준 회 원 : 전남대학교 전산학과 박사과정
†† 정 회 원 : 동신대학교 정보보안학과 교수
††† 정 회 원 : 전남대학교 전자컴퓨터공학부 교수
논문접수 : 2013년 4월 22일
수 정 일 : 1차 2013년 6월 11일
심사완료 : 2013년 6월 26일
* Corresponding Author : Hyeong-Seok Lim(hslim@chonnam.ac.kr)

최근 들어 단일 유전마커에 의한 질병 관련 유전자 탐색 연구에서 여러 유전마커를 한꺼번에 다루어 유전자와 질병과의 연쇄 및 연관성을 분석하는 연구로 점차 옮겨가고 있는 추세이다. 이는 염기서열상에서 가까운 거리에 있는 유전 마커들이 강하게 연쇄되어 있어 단일 유전마커에 대한 연구보다는 더 많은 유전정보를 담아 낼 수 있고 통계량의 검정력도 향상시킬 수 있기 때문이다. 이와 같이 여러 유전마커를 한꺼번에 다루는 연구 중 가장 활발히 연구되고 있는 분야가 일배체형에 대한 연구이다.

일배체형(haplotype)은 사람과 같은 이배체형 생물에서 부계 또는 모계 쪽으로부터 물려받은 하나의 염색체 위의 인접한 유전자좌에 존재하여 함께 전달되는 대립유전자 집합을 말한다. 즉, 한쪽 염색체상에 나타나는 일련의 SNP 서열을 지칭한다[2]. 이배체형 생물의 염색체는 한 쌍의 염색체로 구성되어 있으므로 동일한 SNP 서열 위치에 두 개의 일배체형이 존재한다. 유전자형(genotype)은 부계 쪽 일배체형과 모계 쪽 일배체형의 같은 자리에 있는 두 대립형질의 조합을 나타낸다. 즉, 한 쌍의 일배체형이 하나로 합쳐져 혼합된 서열을 말한다. 일배체형은 개별 SNP 또는 유전자형보다 유전질병 연관성과 관련해 보다 많은 정보를 제공한다. 그러나 실험적인 방법으로 일배체형을 결정하는 작업은 유전자형 서열을 해독하는 방법에 비해 기술적인 제약이 크고 많은 비용을 초래한다[3]. 따라서 일배체형을 결정하기 위한 계산적인 방법이 요구된다.

일배체형을 결정하기 위한 계산적인 방법은 크게 두 종류의 문제를 다루고 있는데, 특정 집단의 유전자형 집합으로부터 일배체형 집합을 도출해내는 일배체형 추론문제(haplotype inference problem)가 그 하나이고, 한 개체로부터 얻은 여러 개의 SNP 단편들로부터 한 쌍의 일배체형을 결정하는 일배체형 조합문제(haplotype assembly problem)가 다른 하나이다. 일배체형 추론문제에 대한 개괄은 [4, 5]를, 일배체형 조합문제에 대한 개괄은 [6, 7]을 참조하기 바란다. 본 논문을 일배체형 조합문제를 다룬다.

DNA 판독 기술의 기술적 한계 때문에 판독된 서열은 오류(error)와 손실(missing or gap)을 포함한다. 따라서 판독된 SNP 단편들 또한 오류와 손실을 포함한다. 그리고 각 단편들은 한 쌍의 염색체로부터 얻어지기 때문에 이들을 자신이 실제로 속하는 염색체와 연계시키는 일은 기술적으로 쉽지 않고 비용도 많이 든다. 일배체형 조합문제란 이러한 손실과 오류가 존재하는 SNP 단편들을 서로 연관이 높은 두 집합으로 나누고, 이로부터 한 쌍의 일배체형을 도출해내는 것을 말한다. 일배체형 조합문제는 많은 계산시간을 요구하는 NP-hard 문제임이 밝혀졌다[8].

일배체형 조합문제를 해결하기 위해 SNP의 단편을 제거하거나 SNP의 위치를 제거하는 등 다양한 모델과 알고리즘들이 제안되었다[6, 7, 9]. 이들 모델 중 오류 발생 환경의 현실적 가정 때문에 최소오류수정(MEC; Minimum Error Correction)모델[10]에 대한 관심이 높다. 최소오류수정을 목적 함수로 사용하는 일배체형 조합문제는 단편 내 손실이 없는 경우에도 NP-hard임이 밝혀졌다[11]. 일배체형 조합문제의 MFR(Minimum Fragment Removal) 모델에 대한 최적해를 보장하는 분기한정 알고리즘[12]이 제시된 이후, MEC 모델에 대한 최적해를 보장하는 분기한정 알고리즘이 제시되었다[10]. 또한 MEC 모델에 유전자형 정보를 추가한 MEC/GI 모델에 대한 최적해를 보장하는 알고리즘이 제시되었다[13]. Wang 등[10]이 제시한 분기한정 알고리즘은 단편의 수에 대해 지수적인 시간복잡도를 갖는다. 하지만 최적해를 보장하는 다른 접근 방법인 K-mec 알고리즘[10]과 같은 특별한 서열 조건을 요구하지 않는다.

분기한정 알고리즘이 효율성을 확보하여 실용적으로 사용될 수 있느냐는 해공간을 탐색하는 효율적인 방법을 찾는데 달려있다. 저자들은 지역탐색 알고리즘과 같은 탐색 알고리즘도 MEC 문제에 대해 상당히 좋은 해를 갖는 다는 것을 관찰하였고 Wang 등이 제시한 분기한정 알고리즘으로부터 재귀적 성질을 발견하였다. 이러한 관찰과 통찰을 통해 특별한 조건 없이 MEC 문제를 실용적인 시간 안에 해결할 수 있는 개선된 분기한정 알고리즘 제안하였으면 꿀벌의 염색체를 대상으로 성능을 확인하였다[14].

본 논문에서는 [14]에서 제안된 분기한정 알고리즘을 다른 생명체에 적용 가능한지를 확인하고자 Daly 등[8]에 의해 제공된 실제 사람의 유전자형 데이터를 사용하여 알고리즘의 성능을 분석하였으며, 실험결과로부터 성공적으로 적용 가능함을 확인하였다.

2. 최소오류수정 문제

실험을 통해 한 쌍의 염색체로부터 길이가 n인 m개의 SNP 단편을 얻었다고 하자. 각 SNP는 원형이거나 변이형 또는 손실(염기 판독 실패)일 수 있다. 따라서 SNP 단편은 {0, 1, -}로 구성된 벡터로 나타낼 수 있다. 여기서 '0'과 '1'은 원형과 변이형을, '-'는 손실을 의미한다. 그리고 일배체형 조합문제의 입력으로 사용 될 길이가 n인 m개의 SNP 단편들은 집합은 m×n 행렬 $M=\{f_{ij}\}$으로 표현할 수 있는데, 이를 SNP 단편 행렬이라 한다. 행렬의 각 행은 SNP 단편 f_i에 해당하고 각 열은 단편들의 SNP 위치에 해당한다.

두 SNP 단편간의 거리는 일치하지 않는 SNP 위치의 개수를 말한다. 단, 어느 한 쪽에 손실이 있거나 두 쪽 다 손실인 경우엔 거리 계산에서 해당 위치는 제외된다. 두 SNP 단편 f_i와 f_j 사이의 거리 $D(f_i, f_j)$의 정의는 다음과 같다.

$$D(f_i,f_j)=\sum_{k=1}^{n} d(f_{ik},f_{jk})$$

여기서 $d(f_{ik},f_{jk})$는 두 단편 f_i와 f_j의 특정 위치 k에 있는 SNP들 간의 거리로 다음과 같다.

$$d(f_{ik},f_{jk})=\begin{cases}1, & \text{if } f_{ik}\neq -, f_{jk}\neq -, \text{and} f_{ik}\neq f_{jk}\\ 0, & otherwise\end{cases}$$

만약 $D(f_i,f_j)>0$이면 두 SNP 단편 f_i와 f_j가 다른 염색체에서 복제되었거나 단편들의 판독과정에서 오류가 있다는 것을 의미하고, 이런 경우 두 SNP 단편 f_i와 f_j는 충돌한다(conflict)라고 한다. 반대로 두 SNP 단편 사이의 거리가 0이면 두 단편은 동일한 일배체형으로부터 복제된 단편임을 의미하고, 이런 경우 두 단편이 양립한다(compatible)라고 한다. 주어진 SNP 단편들이 서로소인 두 집합으로 분리되고 집합 내 모든 단편들 간에 충돌이 없다면 이로부터 한 쌍의 일배체형을 바로 도출할 수 있다. 이렇게 SNP 단편들의 행렬 M의 각 행들의 집합이 충돌이 없는 두 개의 서로소인 집합으로 분할(partition)되면 행렬 M이 타당하다(feasible)라고 한다.

일배체형 조합문제는 주어진 SNP 단편들을 집합 내 단편들간의 충돌이 최소가 되도록 두 집합으로 분할하고, 이를 기반으로 한 쌍의 일배체형을 구하는 것이다. MEC 모델은 MLF(Minimum Letter Flips) 모델이라고도 불리는데, 분할된 집합 내의 충돌을 없애기 위해 단편들의 개별 SNP 값을 변경(flip)시키는 경우를 최소화하는 방법이다. 즉, 주어진 SNP 행렬이 타당해지도록 행렬에서 최소수의 원소들을 변경하는(오류를 수정하는) 방법이다.

주어진 SNP 단편 행렬 M의 각 열들의 집합을 서로소인 두 부분집합 C_0와 C_1로 분할하였다고 하자. 부분집합 C_i(여기서 i는 0 또는 1)의 모든 SNP 단편들이 상호간에 출동이 없게 만들려면 C_i에 속한 SNP 단편의 특정 위치의 SNP 값을 SNP 유형의 개수가 적은 것에서 개수가 많은 것으로 변경(flip)하여야 하는데, 이에 필요한 오류수정수는 다음과 같다.

$$\min(N_{0j}(C_i),N_{1j}(C_i))$$

여기서 $N_{0j}(C_i)$는 분할 C_i에 대한 j번째 행의 0의 개수를 $N_{1j}(C_i)$는 1의 개수를 의미한다. 따라서 분할 $P=(C_0, C_1)$로부터 한 쌍의 일배체형을 조합하는데 필요한 오류수정 수 $E(P)$는 다음과 같음을 알 수 있다.

$$E(P)=\sum_{j=1}^{n}\min(N_{0j}(C_0),N_{1j}(C_0))+\sum_{j=1}^{n}\min(N_{0j}(C_1),N_{1j}(C_1))$$

MEC 문제는 이러한 오류수정의 개수가 최소인 분할 P^*를 찾는 것이다. MEC 모델은 손실이 없는 경우에도 NP-hard이고[11] 1개의 손실이 있는 경우엔 APX-hard임[15]이 밝혀졌다. 현재 이 문제를 해결하기 위해 유전자 알고리즘을 포함한 많은 근사(휴리스틱) 알고리즘과 최적해를 찾는 분기한정 방법이 제시되었다[6, 7].

3. 알고리즘

논문 [14]에서 제시된 개선된 분기한정 알고리즘은 최적해에 근접한 초기 상한값을 구하기 위해 휴리스틱 알고리즘으로 지역탐색 알고리즘을 사용한다. 그리고 개선된 분기한정 알고리즘은 부분 단편들을 대상으로 최적해를 구하고, 원래 문제의 한정함수에서 이를 사용한다. 이러한 방법들을 통해 분기한정 알고리즘의 불필요한 탐색 영역을 현저히 줄임으로써 단편들이 많은 경우에도 분기한정 알고리즘을 적용 가능함을 보였다. 알고리즘의 상세한 내용은 논문 [14]를 참고하기 바란다.

3.1 지역탐색 알고리즘을 사용한 초기 상한 계산

다음 절에 제시될 분기한정 알고리즘의 초기 상한을 계산하기 위해 휴리스틱 방법으로 지역탐색 알고리즘을 사용하였다. m개의 SNP 단편이 주어졌을 경우 이를 두 개의 부분집합으로 나누는 가지수는 2^{m-1}이다. 지역탐색 알고리즘은 2^{m-1}개의 해집합을 대상으로 특정 조건을 만족하는 동안 탐욕적으로 최적해를 탐색해간다. 물론 지역탐색 알고리즘이 최적해를 보장해 주지는 못한다. 지역탐색 알고리즘의 의사코드는 아래와 같다.

알고리즘: LocalSearchForBound
입력: SNP 단편 행렬 M
출력: M의 분할 p, M의 분할 p에 대한 오류수정수 e

```
1  begin
2    if M is feasible then set p as a partition that make M feasible; return (p, 0)
3    Randomly assign each SNP fragment in M into C0 or C1
4    p ← P(C0,C1)
5    e ← E(p)
6    loop
7    |   modified ← FALSE
8    |   foreach p' in Adj(p) do
9    |   |   if e > E(p') then
10   |   |   |   p ← p'
11   |   |   |   e ← E(p')
12   |   |   |   modified ← TRUE
13   |   |   |   break
14   |   |   end if
15   |   end foreach
16   |   if not modified then return (p, e)
17   end loop
18 end
```

Fig. 1. A local search algorithm for the initial upper bound

Fig. 1의 LocalSearchForBound 알고리즘에서 주어진 SNP 단편 행렬이 타당한 경우 일배체형을 결정하는데 필요한 오류수정 수가 0이므로 바로 탐색을 종료한다(라인 2). 주어진 SNP 단편 행렬이 타당하지 않은 경우 임의로 SNP 단편들을 분할하고 이에 대한 일배체형 결정을 위한 최소변경수, $E(p)$를 구한다(라인 3~5). 그런 다음 인접해들을 대상으로 일배체형 결정을 위한 최소변경수를 구하고, 만약 최소변경수를 줄이는 인접해가 존재하면 그 인접해를 기준으로 삼는다(라인 8~15). 새로운 기준에 대해 이와 같은 과정을 반복하여 더 이상 개선된 해가 없을 때까지 계속해서

인접해를 탐색해간다(라인 6~17). 여기서 인접해란 주어진 분할에 대해 SNP 단편 하나가 변경된 분할을 말한다. 예를 들어 3개의 SNP 단편에 대해 분할 $P(\{f_1, f_3\}, \{f_2\})$의 인접해는 $P(\{f_3\}, \{f_1, f_2\})$, $P(\{f_1, f_2, f_3\}, \varnothing)$ 그리고 $P(\{f_1\}, \{f_2, f_3\})$이다. 위의 알고리즘에서 Adj(p)는 분할 p의 모든 인접해를 구하는 함수이다.

3.2 개선된 분기한정 알고리즘

분기한정 알고리즘은 최적화 문제를 해결하기 위한 알고리즘의 설계 방법 중 하나로 해공간 중 일부 영역의 탐색을 회피함으로써 검색의 속도를 증가시키려는 것이다. 이러한 탐색회피가 가능한 이유는 회피 영역을 탐색하기 전의 시점에서 그 영역의 해들이 결코 최적의 해를 갖지 못한다는 사실을 알 수 있기 때문이다. 분기한정 알고리즘의 효율성은 초기 한계값이나 해공간의 탐색방법 등에 영향을 받으며, 가장 중요하게는 불필요한 해공간을 제거하는데 사용되는 한정함수의 영향을 크게 받는다.

일배체형 조합문제를 해결하기 위한 MEC 모델에서 SNP 단편들 일부만을 대상으로 구한 최소변경수는 그 일부를 부분 분할로 갖는 모든 분할들에 대한 최소변경수 보다 작다. 따라서 해공간을 탐색함에 있어 현재까지의 전역 최소변경수보다 단편들 일부에 대한 최소변경수가 크거나 같다면 나머지 단편들을 포함한 모든 분할들의 최소변경수는 당연히 전역 최소변경수보다 크거나 같다. 따라서 이들 분할에 해당하는 해공간은 탐색에서 제외해도 된다. 여기서 전역 최소변경수란 현재까지 탐색한 해공간 전체에서 구한 최소변경수를 말한다.

m개의 SNP 단편이 주어지고 이를 두 개의 집합으로 분할한 $P(C_0, C_1)$에 대하여 단편 f_i가 C_0에 속하면 c_i는 0, 단편 f_i가 C_1에 속하면 c_i는 1이라 하자. 분할 P는 $(c_1, c_2, ..., c_m)$으로 나타낼 수 있다. 이와 같은 방식으로 MEC 모델에서 최소변경수 $E(P)$를 $E(c_1, c_2, ..., c_m)$으로 표시하기로 한다. MEC 문제에 대한 분기한정 알고리즘을 제시한 [10]에서 다음과 같은 정리가 유도되었다.

$$E(c_1, c_2, ..., c_m) \geq E(c_1, c_2, ..., c_k) + E^*(c_{k+1}, c_{k+2}, ..., c_m)$$

여기서 $E(c_1, c_2, ..., c_m)$는 m개의 단편들로 구성된 분할의 최소변경수이고 $E(c_1, c_2, ..., c_k)$는 이중 앞부분 k개로 구성된 분할의 최소변경수이다. $E^*(c_{k+1}, c_{k+2}, ..., c_m)$는 단편 f_{k+1}에서 f_m까지의 $m-k$개로 구성된 단편들의 최적 분할에 대한 최소변경수를 나타낸다. 이는 f_1에서 f_k까지 k개의 단편들로 구성된 분할의 최소변경수와 나머지 단편 f_{k+1}에서 f_m까지의 $m-k$개로 구성된 단편들의 최적 분할에 대한 최소변경수의 합이 전역 최소변경수 보다 크거나 같다면 이들 분할에 해당하는 해공간을 탐색에서 제외해도 된다는 뜻이다. 결과적으로 탐색의 범위를 크게 줄일 수 있다. 하지만 k가 작아, 즉 $m-k$가 커, 최적분할을 구해야 하는 단편들의 수가 크면 그 자체의 해를 구하는데 많은 시간이 요구되므로 $m-k$가 작은 경우에 대해서만 적용할 수 있는 한계가 있다.

논문 [14]에서는 [10]에서 제안한 아이디어를 분기한정 알고리즘의 한정함수에 적용하였으며, 재귀적 성질을 이용하고 룩업테이블(lookup table)을 사용해 알고리즘의 효율성을 향상시킨 개선된 분기한정 알고리즘을 제시하였다. 제안된 알고리즘의 의사코드는 아래와 같다.

알고리즘: ImprovedBranchAndBound
입력: $m \times n$ SNP 단편 행렬 M, 수준 S
출력: 일배체형쌍 (h_0, h_1)

```
1  begin
2      // Phase 1: 룩업테이블 구축
3      E*(c_{m-1}, c_m) ← min(E(c_{m-1}=0, c_m=0),E(c_{m-1}=0, c_m=1))
4      for k=m-2 down to m-S+1 do
5      |   evaluate E*(c_k, ..., c_m) and save it to the lookup table.
6      end for

7      // Phase 2: 원래 크기의 문제에 적용
8      (P, bound) ← LocalSearchForBound(M)
9      RecursiveB&B(M, S)

10     // Phase 3: 분할 P로부터 일배체형 조합
11     (h_0, h_1) ← Assemble(P)
12 end

13 function RecursiveB&B(M, k)
14 begin
15     if k=m then
16     |   l ← E(c_1,⋯,c_{m-1},0)
17     |   r ← E(c_1,⋯,c_{m-1},1)
18     |   if l ≤ r and l < bound then (P, bound) ← ((c_1,⋯,c_{m-1},0), l)
19     |   if l > r and r < bound then (P, bound) ← ((c_1,⋯,c_{m-1},1), r)
20     else if S≤k≤m−1 then
21     |   if E(c_1,⋯,c_{k-1},0)+E*(c_{k+1},⋯,c_m) < bound then RecursiveB&B(M, k+1)
22     |   if E(c_1,⋯,c_{k-1},1)+E*(c_{k+1},⋯,c_m) < bound then RecursiveB&B(M, k+1)
23     else if k<S then
24     |   if E(c_1,⋯,c_{k-2},0) < bound then RecursiveB&B(M, k+1)
25     |   if E(c_1,⋯,c_{k-2},1) < bound then RecursiveB&B(M, k+1)
26     end if
27 end
```

Fig. 2. Improved branch and bound algorithm

Fig. 2의 개선된 분기한정 알고리즘에서는 수준 S를 사용하여 함수 RecursiveB&B에서 k가 수준 S보다 크거나 같은 경우(라인 20) 하한을 구할 때 $E(c_1, c_2, ..., c_k) + E^*(c_{k+1}, c_{k+2}, ..., c_m)$를 사용하고 k가 수준 S보다 작은 경우(라인 23)에는 $E(c_1, c_2, ..., c_k)$를 사용한다. 이렇게 구한 하한값이 전역 최소변경수(bound)보다 작은 경우에만 재귀적으로 탐색을 지속하며 최소변경수보다 큰 경우에는 재귀적 탐색을 중지한다(라인 20~25). k가 m과 같은 경우(라인 15) 즉, 마지막 한 개의 SNP 단편을 제외한 모든 SNP 단편들의 대한 분할이 결정된 경우에는 마지막 단편을 부분집합 C_0와 C_1에

가각 귀속시켜 변경수를 구한다(라인 16~17). 이렇게 구한 변경수가 전역 최소변경수보다 작은 경우 해당 분할을 저장하고 전역 최소변경수를 수정한다(라인 18~19). 초기 상한 즉, 전역 최소변경수의 초기값은 Fig. 1의 알고리즘 LocalSearchForBound에서 구해진 변경수를 사용한다(라인 8).

본 알고리즘에서는 주어진 수준 S보다 크거나 같은 k에 대해 $E^*(c_k, c_{k+1}, ..., c_m)$을 사전에 계산하여 룩업테이블을 구축하여 사용한다(라인 2~5). $E^*(c_m)$은 당연히 0이다. $E^*(c_{m-1}, c_m)$은 두 가지의 가능한 분할(둘 모두 C_0 또는 C_1에 속하거나 하나는 C_0에 나머지 하나는 C_1에 속한 경우)에 대해 최소변경수를 계산하여 이를 비교하여 최적값을 구한다(라인 3). k=m-2부터 k=m-S+1이 될 때까지 이전에 구해 놓은 $E^*(c_{k+1}, ..., c_m)$을 활용하여 최적값을 구한다(라인 4~6). 최종적으로 구해진 분할로부터 각 부분집합의 공통 일배체값을 계산하여 일배체형 조합을 구한다(라인 11).

4. 실험 결과 및 분석

개선된 분기한정 알고리즘은 C언어로 구현하였으며 32비트 팬티엄-4 PC(2.8GHz CPU, 1GB RAM)에서 실험하였다.

4.1 실험 데이터

인간 유전체 데이터 셋은 Daly 등[8]에 의해 제공된 IBD5 데이터를 이용하였다. 이 데이터는 사람의 부-모-자식으로부터 염색체 5q31의 500 Kb 영역을 대상으로 103 개의 유전자형들을 결정한 것이다. 다른 논문들[9, 10]과 동일한 방법으로 총 258 쌍의 일배체형을 도출하고 이 중 손실률이 20% 이상인 것을 제거하여 147 개의 쌍으로 구성된 테스트 집합을 실험자료로 삼았다. 현재 공개된 SNP 단편들에 대한 자료는 얻을 수 없었으므로 147 개의 일배체형 쌍에 대해 일정한 손실율과 오류율을 사용하여 30~60 개의 모의 SNP 단편들을 생성하였다. 추가적으로 손실율과 오류율에 따른 알고리즘의 성능평가를 위해 전산학적으로 합성된 데이터 셋에 대하여 실험을 실시하였다.

4.2 성능비교

본 절에서는 Wang 등[10]에 의해 제시된 분기한정 알고리즘과 개선된 분기한정 알고리즘간의 성능을 비교한다.

Fig. 3은 Daly 데이터 집합으로부터 얻은 147개의 일배체형 쌍으로부터 각 매개변수를 단편의 개수 m=30, 손실율 GR=0.1, 오류율 ER=0.2로 고정하고 생성한 실험 사례들을 대상으로 두 가지 알고리즘을 적용한 경우의 소요시간을 비교한 것이다. 개선된 분기한정 알고리즘에서 수준 S=10을 적용하였다. Wang 등[10]에 의해 제안된 분기한정 알고리즘(Fig. 3a)에 비해 개선된 분기한정 알고리즘(Fig. 3b)의 효율성이 크다는 사실을 확인할 수 있다. 개선된 분기한정 알고리즘의 경우 147 개의 실험 사례들 중 24 개의 사례만이

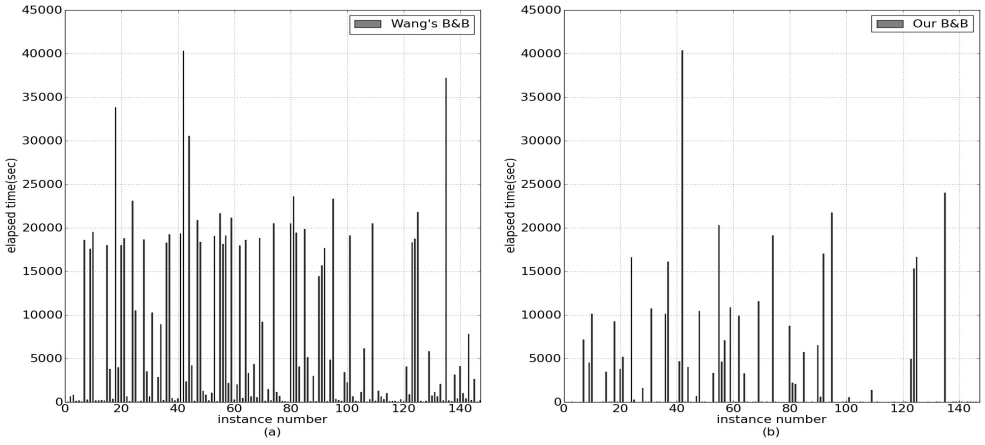

Fig. 3. Run time histograms for Daly data set (a) Wang's branch and bound algorithm (b) Improved branch and bound algorithm.

Table 1. The results of the number of successes for the human data set

$m(S)$	ER = 0.1		ER = 0.2		ER = 0.3	
	Wang's[1)]	Improved[2)]	Wang's	Improved	Wang's	Improved
40(10)	84 (57%)	135 (92%)	6 (4%)	109 (74%)	0 (0%)	56 (38%)
50(10)	51 (35%)	127 (86%)	0 (0%)	107 (73%)	0 (0%)	45 (31%)
60(15)	19 (13%)	123 (84%)	0 (0%)	106 (72%)	0 (0%)	51 (35%)

1) Wang's branch and bound algorithm

2) Improved branch and bound algorithm

Table 2. The results of the average running time for the human data set

	ER = 0.1		*ER* = 0.2		*ER* = 0.3	
m(*S*)	Wang's	Improved	Wang's	Improved	Wang's	Improved
40(10)	275.5	80.2	2604.1	80.4	*N/A*	585.1
50(10)	514.3	54.9	*N/A*	14.6	*N/A*	806.1
60(15)	1849.7	59.3	*N/A*	25.6	*N/A*	813.4

Table 3. The results of the number of successes for the synthesized data set

		ER = 0.05		*ER* = 0.1		*ER* = 0.15		*ER* = 0.2	
GR=0.1	*m*(*S*)	Wang's	Improved	Wang's	Improved	Wang's	Improved	Wang's	Improved
	30(10)	20(100%)	20(100%)	20(100%)	20(100%)	20(100%)	20(100%)	19(95%)	20(100%)
	40(10)	20(100%)	20(100%)	17(85%)	20(100%)	17(85%)	20(100%)	0(0%)	20(100%)
	50(10)	19(95%)	20(100%)	3(15%)	20(100%)	3(15%)	20(100%)	0(0%)	20(100%)
		ER = 0.05		*ER* = 0.1		*ER* = 0.15		*ER* = 0.2	
GR=0.2	*m*(*S*)	Wang's	Improved	Wang's	Improved	Wang's	Improved	Wang's	Improved
	30(10)	20(100%)	20(100%)	20(100%)	20(100%)	20(100%)	20(100%)	19(95%)	20(100%)
	40(10)	20(100%)	20(100%)	18(90%)	20(100%)	18(90%)	20(100%)	0(0%)	20(100%)
	50(10)	19(95%)	20(100%)	8(40%)	19(95%)	8(40%)	19(95%)	0(0%)	18(900%)

Table 4. The results of the average running time for the synthesized data set

		ER = 0.05		*ER* = 0.1		*ER* = 0.15		*ER* = 0.2	
GR=0.1	*m*(*S*)	Wang's	Improved	Wang's	Improved	Wang's	Improved	Wang's	Improved
	30(10)	0.3	0	7.05	0	7.1	0	134.053	0.55
	40(10)	11.6	0	211.412	7.25	211.294	7.25	N/A	2.3
	50(10)	38	0	392.667	0	392.667	0.05	*N/A*	34
		ER = 0.05		*ER* = 0.1		*ER* = 0.15		*ER* = 0.2	
GR=0.2	*m*(*S*)	Wang's	Improved	Wang's	Improved	Wang's	Improved	Wang's	Improved
	30(10)	1.05	0	26.5	0	26.4	0.05	384.684	1.35
	40(10)	48.35	0	191.778	0.05	191.944	0	N/A	74.7
	50(10)	107.684	0.05	596.75	0.053	596.625	0.053	N/A	7.5

5,000 초가 넘었고 100 개 이상의 사례가 60 초 이하에 해결되었다. 이러한 사실은 휴리스틱 알고리즘을 적용해야 할 몇몇 경우를 제외하고는 최적해를 보장하는 개선된 분기한정 알고리즘을 일반적으로 사용할 수 있음을 시사해준다.

Table 1은 Wang 등[10]이 제시한 알고리즘과 본 논문에서 제시된 알고리즘의 성능 비교를 위해 논문 [8]에 제시된 데이터 셋으로부터 생성된 실험사례에 대한 제한시간 내 성공한 횟수를 비교한 것이다. Table 1의 값들은 손실율 *GR*을 0.1로 고정하고 단편의 개수 *m*, 수준 *S*, 오류율 *ER* 등의 매개변수 값을 달리해가며 생성한 147 개의 실험 사례 중 성공한 횟수와 백분율을 나타낸다. 이때 성공한 사례란 3,600초 이내에 최적해를 찾은 사례를 말한다. 개선된 분기한정 알고리즘의 경우 오류율이 낮은 경우 대부분의 사례에서 최적해를 찾고 있음을 알 수 있다. 최근의 서열 판독기의 오류율이 높지 않다는 사실을 감안하면 개선된 알고리즘이 실제 응용에서 허용 가능한 시간 내에 문제를 해결할 가능성이 높음을 알 수 있다. Table 2에서는 제한 시간 3,600초에 대해 최적값을 찾는데 성공한 사례에 대한 평균 실행시간을 비교한 것이다. Table 2에서 볼 수 있듯이 성공한 사례들의 평균 실행 시간은 3,600초에 비해 훨씬 작았다.

추가적으로 손실율과 오류율에 따른 알고리즘의 성능평가를 위해 전산학적으로 합성된 데이터 셋에 대하여 실험을 실시하였다. 합성 데이터 셋은 길이 30인 랜덤 시퀀스를 20쌍(homologous rate는 70%)을 생성하고 단편 개수 *m*은 30, 40, 50으로, 손실률 *GR*은 0.1과 0.2로, 오류율 *ER*은 0.05, 0.1, 0.15, 0.2로 하여 실험하였다. Table 3과 4는 합성된 데이터 셋에 대한 제한시간 내 성공 횟수와 성공한 사례에 대한 평균 실행시간을 비교한 것이다. 실제 인간 유전체의 실험과 마찬가지로 제한시간은 3,600초를 사용하였다. 현재의 염기서열 결정 기술은 낮은 오류율을 보인다. 표에서 보인 바와 같이 개선된 분기한정 알고리즘은 낮은 오류율에 대해 대부분 짧은 시간 내에 해결함을 확인할 수 있다.

5. 결 론

일배체형 조합문제는 생물정보학 등에서 중요한 문제 중 하나이다. 본 논문은 [14]에서 제안된 개선된 분기한정 알고리즘을 다른 일반적인 생명체에 대하여 적용가능성을 확인하고자 Daly 등이 제시한 실제 인간 유전자형을 대상으로 실험을 확장하여 수행하였다. 개선된 분기한정 알고리즘은 최적해에 근접한 초기 상한값을 구하기 위해 지역탐색 알고리즘을 사용하였다. 그리고 개선된 분기한정 알고리즘은 부분 단편들을 대상으로 최적해를 구하고 이를 룩업테이블에 저장하여 원래 문제의 한정함수에 사용하였다.

실험 결과에서 알 수 있듯이 인간 유전자 데이터에 대해 개선된 분기한정 알고리즘을 성공적으로 적용 가능함을 확인하였다. 차세대염기서열분석(NGS) 방법이 제시되면서 적은 비용에 오류율이 낮은 염기서열을 결정하고 있다. 개선된 분기한정 알고리즘은 오류율이 낮을 때 높은 효율성을 보여주고 있다. 따라서 현장에서 제기되는 대부분의 경우에 최적해를 보장하는 개선된 분기한정 알고리즘의 적용 가능성이 충분하다고 여겨진다.

참 고 문 헌

[1] The International HapMap Consortium, "A second generation human haplotype map of over 3.1 million SNPs," Nature, Vol.449, pp.851-861, 2007.

[2] Wikipedia, http://en.wikipedia.org/wiki/Haplotype.

[3] C. Burgtorf, P. Kepper, M. Hoehe, C. Schmitt, R. Reinhardt, H. Lehrach, S. Sauer, "Clone-based systematic haplotyping (CSH): A procedure for physical haplotyping of whole genomes," *Genome Research*, Vol.13, pp.2717-2724, 2003.

[4] A. Graça, I. Lynce, J. Marques-Silva, and A.L. Oliveira, "Haplotype inference by Pure Parsimony: a survey", *Journal of Computational Biology*, Vol.17, pp.969-992, 2010.

[5] S.R. Browning and B.L. Browning, "Haplotype phasing: existing methods and new developments," *Nature Reviews Genetics*, Vol.12, pp.703-714, 2011.

[6] X. S. Zang, R. S. Wang, L. Y. Wu, and L. Chen, "Models and algorithms for haplotyping problem," *Current Bioinformatics*, Vol.1, No.1, pp.105-114, 2006.

[7] R. Schwartz, "Theory and Algorithms for the Haplotype Assembly Problem," *Communication in Information and Systems*, Vol.10, No.1, pp.23-38, 2010.

[8] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, E. S. Lander, "High-resolution haplotype structure in the human genome," *Nature Genetics*, Vol.29, pp.229-232, 2001.

[9] S. H. Kang, I. S. Jeong, H. G. Cho, H. S. Lim, "HapAssembler: A web server for haplotype assembly from SNP fragments using genetic algorithm," *Biochemical and Biophysical Research Communications*, Vol.397, pp.340-344, 2010.

[10] R. S. Wang, L. Y. Wu, Z. P. Li, X. S. Zhang, "Haplotype reconstruction from SNP fragments by minimum error correction," *Bioinformatics*, Vol.21, pp.2456-2462, 2005.

[11] R. Lippert, R. Schwartza, G. Lancia, and S. Istrail, "Algorithmic Strategies for the SNPs Haplotype Assembly Problem," *Briefings in Bioinformatics*, Vol.3, No.1, pp.23-31, 2002.

[12] M, Xie, J. Wang, J. Chen, "A practical exact algorithm for the individual haplotyping problem MEC," in *Proceedings of the 2008 International Conference on BioMedical Engineering and Informatics*, pp.72-76, 2008.

[13] J. Wang, et al., "A Practical Exact Algorithm for the Individual Haplotyping Problem MEC/GI," *Algorithmica*, Vol.56, pp.283-296, 2010.

[14] H. S. Lim, I. S. Jeong, and S. H. Kang, "Individual haplotype assembly of Apis mellifera (honeybee) using a practical branch and bound algorithm," *Journal of Asia Pacific Entomology*, Vol.15, pp.375-381, 2012.

[15] R. Cilibrasi, L. Iersel,S. Kelk, and J. Tromp, "The Complexity of the Single Individual SNP Haplotyping Problem," *Algorithmica*, Vol.49, No.1, pp.13-36, 2007.

최 문 호

e-mail : howork@gmail.com
1993년 전남대학교 전산통계학과(학사)
1995년 전남대학교 전산학과(석사)
2000년~2004년 동신대학교 디지털영상 매체기술혁신센터 연구교수
2000년~현 재 전남대학교 전산학과 박사과정
관심분야 : 생물정보학, 알고리즘, 정보보안, 가상환경 등

강 승 호

e-mail : kinston@gamil.com
1994년 전남대학교 전산학과(학사)
2003년 전남대학교 전산학과(석사)
2009년 전남대학교 전산학과(박사)
2009년~2010년 목포대학교 정보산업 연구소 전문연구원
2010년~2013년 국가수리과학연구소 수리생물학연구팀 연구원
2013년~현 재 동신대학교 정보보안학과 교수
관심분야 : 생물정보학, 정보보안, 알고리즘 등

임 형 석

e-mail : hslim@chonnam.ac.kr

1983년 서울대학교 컴퓨터공학과(학사)

1985년 학국과학기술원 전산학과(석사)

1993년 학국과학기술원 전산학과(박사)

1996년~1997년 미국 Purdue대학 방문과학자

1987년~현 재 전남대학교 전자컴퓨터공학부 교수

관심분야 : 알고리즘, 그래프이론, 생물정보처리 등