신뢰도를 가진 SNP 단편들과 유전자형으로부터 일배체형 조합

(Haplotype Assembly from Weighted SNP Fragments and Related Genotype Information)

강 승 호 † 정 인 선 † 최 문 호 † 임 형 석 ††
(Seung-Ho Kang) (In-Seon Jeong) (Mun-Ho Choi) (Hyeong-Seok Lim)

요 약 Minimum Letter Flips(MLF) 모델과 Weighted Minimum Letter Flips(WMLF) 모델은 일배체형 조합문제(haplotype assembly problem)를 해결하기 위한 모델들이다. 그러나 MLF 모델이나 WMLF 모델은 SNP(Single Nucleotide Polymorphism) 단편들에 손실과 오류가 적은 경우에만 효과적이다. 본 논문은 WMLF모델의 개선을 목적으로 유전자형 정보를 추가한 WMLF/GI 모델과 문제를 제시한다. 새로 제시한 문제가 NP-hard임을 증명하고, 정확성이 높고 효율적인 문제 해결을 위해 유전자 알고리즘을 설계한다. 실험 결과를 통해 새로운 모델이 기존의 모델들에 비해 SNP 단편들에 손실과 오류가 많은 경우에도 높은 정확성을 가짐과 유전자형 정보가 유전자 알고리즘의 수렴속도를 크게 개선함을 보인다.

키워드 : 일배체형 조합, SNP, WMLF, WMLF/GI, 유전자알고리즘

Abstract The Minimum Letter Flips (MLF) model and the Weighted Minimum Letter Flips (WMLF) model are for solving the haplotype assembly problem. But these two models are effective only when the error rate in SNP fragments is low. In this paper, we first establish a new computational model that employs the related genotype information as an improvement of the WMLF model and show its NP-hardness, and then propose an efficient genetic algorithm to solve the haplotype assembly problem. The results of experiments on random data set and a real data set indicate that the introduction of genotype information to the WMLF model is quite effective in improving the reconstruction rate especially when the error rate in SNP fragments is high. And the results also show that genotype information increases the convergence speed of the genetic algorithm.

Key words : haplotype assembly, SNP, WMLF, WMLF/GI, genetic algorithm

1. 서 론

사람의 유전체(genome) 서열이 전부 밝혀짐에 따라 유전적 차이에 대한 연구가 유전학에서 중요한 주제가 되었다[1]. 인간은 염색체(DNA) 수준에서 99%가 동일하고 유전병은 나머지 1% 영역에서의 차이에 기인함을 완성된 유전체 정보로부터 알 수 있다[2]. 인간 유전체에서 유전적 변이를 가장 풍부하게 보여주는 유전 마커(genetic marker)로 SNP이 대표적이며, 이의 이해가 인간의 질병 치료와 약품 설계 그리고 새로운 의료 기구 생산에 대한 능력을 증가시킬 것으로 예측하고 있다.

SNP이란 DNA서열상의 특정 위치에 존재하는 단일 염기의 변이를 말한다. 변이의 범위는 제한되어 있고 각 변이를 대립유전자(allele)라 한다. SNP은 일반적으로 원형(wild type)과 돌연변이형(mutant type)의 두 가지 대립유전자를 갖는데 각각 0과 1로 표기한다. 그리고 그림 1처럼 특정 염색체의 SNP 서열을 일배체형이라 한다. 인간과 같은 이배체(diploid) 생물은 유전체가 한 쌍의 염색

·이 논문은 2005년도 전남대학교 연구년 교수연구비 지원에 의하여 연구되었음.

† 학생회원 : 전남대학교 전산학과
kinston@gmail.com
isjung0@hotmail.com
howork@paran.com
†† 종신회원 : 전남대학교 전자컴퓨터공학부 교수
hslim@chonnam.ac.kr

논문접수 : 2008년 8월 6일
심사완료 : 2008년 10월 17일

정보과학회논문지: 시스템 및 이론 제35권 제11호(2008.12)

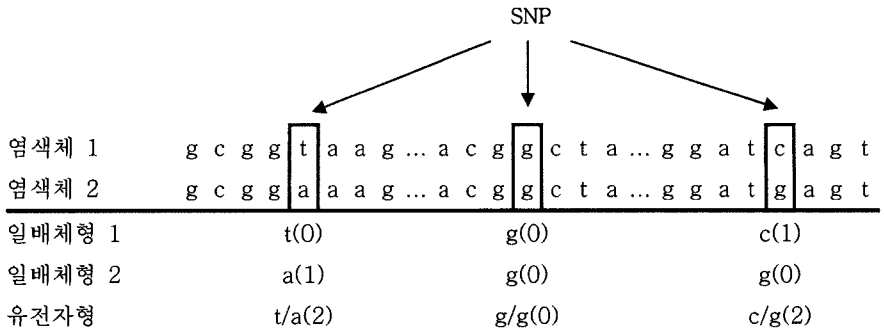

그림 1 한 쌍의 일배체형과 유전자형 예

체로 구성되어있기 때문에 두 개의 일배체형이 존재한다. 유전자형(genotype)이란 상동염색체에 대한 두 일배체형의 조합(conflation)을 말한다. SNP에 대해 두 대립유전자형이 동일하면 이 SNP 위치(site)를 동형(homozygous)이라 하고 대립유전자형에 따라 0 또는 1로 나타내고, 서로 다르면 이형(heterozygous)이라 하며 2로 나타낸다. 질병 연구에서는 일배체형이 유전자형보다 중요한 역할을 한다[3]. 하지만 현재의 SNP 판독 기술은 SNP의 위치와 대립유전자형을 알아낼 수 있을 뿐 대립유전자가 두 염색체 중 어떤 염색체의 것인지를 판별하지는 못한다. 따라서 일배체형을 결정하는 문제가 유전자형을 결정하는 일보다 훨씬 어렵다. 이러한 어려움을 극복하기 위해 전산학적 관점에서 두 부류의 문제가 정의 되었다. 특정 집단의 유전자형 집합으로부터 일배체형 집합을 추론하는 일배체형 추론문제(haplotype inference problem)가 그 하나이고 한 사람으로부터 얻은 여러 개의 SNP 단편들로부터 한 쌍의 일배체형을 조합해내는 일배체형 조합문제가 다른 하나이다. 본 논문은 일배체형 조합문제를 다룬다. 일배체형 조합문제는 손실(missing or gap)과 오류(error)가 존재하는 SNP 단편들을 두 부분으로 나누고 이로부터 한 쌍의 일배체형을 결정하는 것이다.

일배체형 조합문제에는 여러 가지 모델이 제시되어 있는데 각각 다른 실험상의 문제들을 가정하고 있다[4]. 이중 Minimum Letter Flips(MLF) 모델과 이 모델의 가중치 버전인 Weighted Minimum Letter Flips (WMLF) 모델은 하나의 생물 개체로부터 모든 단편들이 얻어지고 단편들의 SNP 판독에 손실과 오류가 있다고 가정한다. 여기서 손실이란 SNP을 판독하지 못한 경우를 말하고 오류란 잘못 판독한 경우를 말한다. WMLF 모델은 이러한 가정에 각 SNP의 염기를 판독하는데 신뢰도라는 가중치를 추가한다[5]. 이 가중치는 SNP 판독 기계가 판독된 염기에 대해 가지는 정확성을 나타낸다. 두 모델이 제시한 문제들은 단편들에 손실이 없는 경우에도 NP-hard임이 증명되었다[6-8]. 그리고 Zhao 등[8]은 WMLF 모델이 MLF 모델보다 일배체형을 조합하는데 높은 정확성을 가짐을 보여주었다. 그러나 두 모델은 SNP 판독상의 손실률과 오류율이 낮은 경우에 효과적이다. MLF 모델[1]의 정확성을 향상시키기 위해 유전자형을 도입하는 아이디어가 제시되었다. 유전자형은 적은 비용으로 얻을 수 있는 정보이기 때문에 현실적으로 중요한 전략이다. 이후 유전자형을 MLF 모델에 도입한 방법들이 제시되었다[1,9,10]. 그러나 일배체형 결정문제에 보다 정확한 WMLF 모델에 유전자형 정보를 도입한 방법이 아직까지 제시되지 않았다. 본 논문에서는 WMLF 모델에 유전자형을 도입한 새로운 WMLF/GI 모델과 이를 해결하기 위한 유전자 알고리즘을 제시한다. 그리고 유전자형의 도입이 조합의 정확성뿐 아니라 유전자 알고리즘의 빠른 수렴에도 도움이 된다는 사실을 보인다.

본 논문은 다른 자료들을 대상으로 저자들이 이전에 제시했던 모델과 알고리즘[11]의 실험을 확장하고 MLF/GI 모델과의 비교도 추가한 것이다. 논문의 구성은 다음과 같다. 2장에서는 문제에 대한 정의와 문제의 복잡도를 보이고, 3장에서는 제시된 문제를 해결하기 위한 알고리즘을 설계한다. 마지막으로 4장과 5장에서 실험결과에 대해 분석하고 결론을 맺는다.

2. 문제 정의

문제의 정의에 필요한 용어들을 먼저 정의한다. 실험을 통해 한 쌍의 염색체로부터 길이가 n인 m개의 SNP 단편들을 얻었다고 하자. 각 SNP는 원형이거나 돌연변이형, 즉 0과 1값을 갖는다. 이러한 단편들은 그림 2처럼 {0, 1, -}로 구성된 $m \times n$ 행렬 M으로 표현되는데 이를 SNP 행렬이라 부른다. '-' 기호는 손실을 나타낸다. 행렬의 각 행은 SNP 단편 f_i에 해당하고 각 열은 단편들의 SNP 위치에 해당한다. SNP 판독기계는 이러한 단편들의 SNP 값에 신뢰도를 부여하는데 이는 SNP 값이 올바로 판독 되었는지에 대한 0과 1사이의 확률

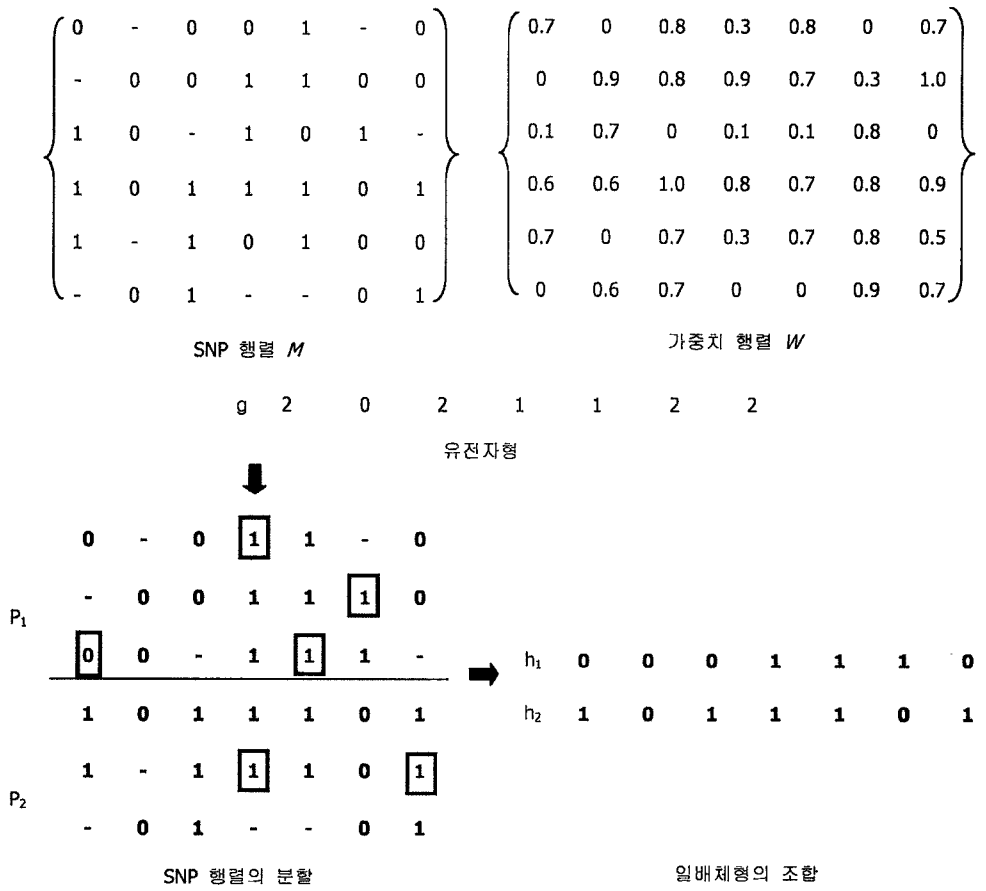

그림 2 WMLF/GI 문제의 예

값으로 나타낸다. SNP 값에 대한 신뢰도는 $m \times n$의 가중치 행렬 W로 표현되고 행렬의 원소 w_{ij}는 행렬 M의 원소인 SNP 위치 f_{ij}의 값에 대한 신뢰도를 나타낸다. f_{ij}가 '–' 값을 가지면 신뢰도를 0으로 한다. 서로 다른 값을 가진 SNP 위치는 가중치가 낮은 값을 다른 값으로 바꾸면 적은 비용으로 일치시킬 수 있으므로 두 단편 f_i와 f_j의 SNP 위치 사이의 거리는 그들의 가중치를 사용하여 다음과 같이 정의한다.

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

만약 한쪽이 SNP 단편이 아니고 일배체형 h_j인 경우엔 SNP 단편의 가중치를 사용한다. 즉, $d(f_{ik}, h_{jk}) = w_{ik}$이다. 단편 f_i과 f_j 사이의 거리는 두 단편의 SNP 전체를 일치시키는데 드는 최소 가중치의 합으로 정의 한다.

$$D(f_i, f_j) = \sum_{k=1}^{n} d(f_{ik}, f_{jk}) \quad (2)$$

만약 $D(f_i, f_j)>0$이면 두 단편 f_i과 f_j가 다른 염색체에서 복제되었거나 SNP 판독에 오류가 있었음을 의미하고 이런 경우를 충돌한다(conflicting)라고 한다. 마찬가지로 일배체형 h_j와 SNP 단편 f_j 사이의 거리는 아래와 같이 정의한다.

$$D(h_i, f_j) = \sum_{k=1}^{n} d(h_{ik}, f_{jk}) \quad (3)$$

모든 SNP 단편들이 서로소인 두 집합으로 분리되고 집합내의 모든 단편들 간에 충돌이 없으면 SNP 행렬을 타당하다(feasible)라고 한다.

그리고 그림 1처럼 유전자형 $g=(g_1, g_2, ..., g_n)$에 대해 i번째 SNP 위치가 모두 원형의 대립유전자를 가지면 0을 g_i에 부여하고 돌연변이형의 대립유전자를 가지면 1을 부여한다. 만약 SNP 위치가 이형이면 2를 부여한다. 그리고 한 쌍의 일배체형 h_1과 h_2가 각 SNP 위치에 대해서 아래의 조건을 만족하면 이 일배체형 쌍은 유전자형과 양립한다(compatible)라고 한다.

$$\begin{cases} \text{if } g_k \neq 2, \ h_{1k} = h_{2k} = g_k \\ \text{if } g_k = 2, \ h_{1k} = 0, h_{2k} = 1 \text{ or } h_{1k} = 1, h_{2k} = 0 \end{cases} \quad (4)$$

MLF 문제는 유전자형과 가중치 행렬을 사용하지 않고 단지 주어진 SNP 행렬로부터 행렬의 원소 값들을 최소 개수로 변경하여 두 집합이 타당하도록 분할 하고 두 일배체형을 결정한다. WMLF 문제는 MLF 문제에 가중치 행렬만을 추가한 것이고 MLF/GI 문제는 MLF 문제에 유전자형을 추가하여 유전자형과의 양립까지 요구한다. 이에 대해 새로운 모델인 WMLF/GI 문제는 다음과 같이 정의한다.

정의 1. WMLF/GI 문제 SNP 행렬 M과 가중치 행렬 W 그리고 유전자형 g가 주어지면, "가중치의 합이 최소"이면서 변경 후의 SNP 행렬이 "타당"하고 유전자형과 "양립"하도록 SNP 행렬의 원소 값들을 0에서 1로 혹은 그 반대로 변경하라. 즉, SNP 단편들을 최소의 가중치로 개별 원소들을 변경하여 집합내의 단편들끼리 상호 충돌이 없는 서로 소인 두 집합으로 분리하고 유전자형과 양립하도록 한 쌍의 일배체형을 결정하라.

그림 2의 예처럼 두 개의 행렬과 유전자형이 주어지면 SNP 위치의 유전자형을 고려하면서 가중치의 합이 최소가 되도록 SNP 값들을 변경하여 서로소인 두 집합으로 SNP 단편들을 분리하고 이 두 집합으로부터 한 쌍의 일배체형을 결정한다. 예제에서 SNP 값의 변경에 따르는 가중치 값의 합은 1.6이며, 이는 유전자형과 양립하도록 SNP 행렬을 이분하는 어떤 변경보다도 작은 값이다.

정리 1. WMLF/GI 문제는 NP-hard 문제이다.

증명. MLF 모델에 유전자형 정보를 도입한 MLF/GI 문제는 Zhang 등에 의해 NP-hard임이 증명되었다[10]. WMLF/GI 문제가 NP-hard 문제임을 증명하기 위해 MLF/GI 문제를 WMLF/GI 문제로 바꿀(reduction) 것이다.

MLF/GI 문제의 임의의 사례(instance) 하나를 단지 특별한 가중치 행렬 W^*를 첨가함으로써 WMLF/GI 문제의 사례로 바꿀 수 있다. SNP 행렬의 원소 값이 0이거나 1을 갖는 경우, 즉 손실이 아닌 경우에는 대응하는 가중치 행렬 W^*의 원소에 1을 부여하고 SNP 행렬의 원소 값이 '-'인 경우엔 0을 부여한다.

MLF/GI 문제에서 SNP 단편 간의 거리나 단편과 일배체형 간의 거리는 해밍거리 즉, 손실을 제외한 서로 다른 대립유전자를 가진 위치의 개수이다. 그런데 W^*의 모든 원소들은 손실을 제외하고 모두 1이므로 WMLF/GI 문제의 서로 다른 대립유전자의 거리도 1이 되어 거리 계산식 (1), (2), (3)은 MLF/GI 문제의 거리 계산식과 동일하게 된다. 따라서 두 문제는 상호 필요충분조건이 성립한다. MLF/GI 문제는 WMLF/GI 문제의 가중치 행렬 W^*을 가진, 즉 손실을 제외한 SNP 판독에 오류가 있더라도 언제나 100%의 신뢰를 갖는 특별한 경우이다.

그리고 문제를 바꾸는 과정엔 다항시간(polynomial time)이 소요된다는 것을 쉽게 알 수 있다. □

3. WMLF/GI 문제를 해결하기 위한 유전자 알고리즘의 설계

이 장에서는 WMLF/GI 문제를 해결하기 위해 유전자 알고리즘[12]을 제시한다. 유전자 알고리즘은 유용한 메타 휴리스틱 알고리즘으로 알려져 있으며 전산생물학을 포함한 여러 분야에서 성공적으로 사용되고 있다.

3.1 가설 공간

유전자 알고리즘에서 한 세대내의 각 개체들은 비트 벡터로 표현된다. 그리고 이 개체는 SNP 단편들의 분할을 나타내며 WMLF/GI 문제의 가능한(feasible) 해 중 하나이다. 가설 공간상의 개체 하나의 길이는 SNP 단편들의 개수이다. 개체의 i번째 위치의 값 0, 1은 i번째 단편의 분할 소속을 나타낸다. 길이 m인 가능한 모든 비트 벡터들이 가설 공간을 구성하고 아래와 같이 표현한다.

$$H = \{(f_1, f_2, ..., f_m) \mid f_i \in \{0,1\}, i = 1,2,...,m\} \quad (5)$$

가설 공간의 크기는 2^{m-1}이다.

3.2 유전자형 정보를 이용한 초기 세대 생성

초기 세대의 개별 개체들은 각 원소 f_i에 0 또는 1을 임의로 부여하여 생성할 수 있다. 하지만 생성과정에 유전자형 정보를 사용하면 유전자 알고리즘의 효율성을 개선할 수 있다.

만약 g_i가 0 이거나 1이면 이는 SNP 단편들의 i번째 위치가 모두 동일하게 0 이나 1 값을 가져야 한다는 것을 의미하는데, 이에 위배된 값들을 사전에 수정 할 수 있다. 그러나 g_i가 2이면 이 위치의 값은 사전에 결정할 수 없다. 이러한 사정에 따라 우선 유전자형이 0 또는 1을 갖는 위치의 모든 단편들을 대응하는 값으로 수정하면 새로운 SNP 행렬 M'를 얻는다. 새로 얻은 행렬 M'에서 두 단편 f_i과 f_j가 $D(f_i, f_j) = 0$이면 동일한 염색체에서 복제된 단편으로 여길 수 있으므로 충돌이 없는 모든 단편들에게 동일한 값을 부여한다. 이렇게 하면 가설 공간은 줄어들고 정확성을 유지하면서도 유전자 알고리즘의 수렴속도가 빨라진다.

3.3 분할로부터 일배체형을 조합해내기 위한 규칙

한 개체의 분할 $P = (P_1, P_2)$로부터 유전자형 정보의 지도아래 한 쌍의 일배체형을 결정하는 방법을 설명한다. $C_{1j}(P_l)$와 $C_{1j}(P_l)$는 부분 P_l의 j번째 열에 대한 원형

과 돌연변이형의 가중치 합을 각각 나타낸다. 즉, $C_{0j}(P_l)=\sum_{f_i\in P_l, f_{ij}=0} w_{ij}$, $C_{1j}(P_l)=\sum_{f_i\in P_l, f_{ij}=1} w_{ij}$ 이다. 만약 $g_j \neq 2$이면 두 일배체형의 j번째 위치의 값은 가중치 합에 상관없이 $h_{1j}=h_{2j}=g_j$ 으로 한다. $g_j=2$ 이면 h_{1j}, h_{2j} 는 아래의 수식에 의해 결정한다.

$$h_{lj}=\begin{cases}0, & \text{if } C_{0j}(P_l)>C_{1j}(P_l)\\ 1, & \text{otherwise}\end{cases} \quad (6)$$

여기서 $l=1,2$ 이고, $j=1,2,\cdots,n$ 이다. 이때 $h_{1j}\neq h_{2j}$ 이면 h_1 과 h_2 은 j 번째 위치에서 유전자형 g 와 양립하지만 이 두 값이 같으면 g 와 양립할 수 없어 h_1 과 h_2 을 수정해야 한다. 두 값이 $h_{1j}=h_{2j}=1$인 경우에는 $C_{1j}(P_1)$-$C_{0j}(P_1)$와 $C_{1j}(P_2)$-$C_{0j}(P_2)$을 비교하여 후자가 더 큰 경우엔 $h_{1j}=0$로 하고 이외의 경우엔 $h_{2j}=0$로 한다. $h_{1j}=h_{2j}=0$인 경우에는 $C_{0j}(P_1)$-$C_{1j}(P_1)$와 $C_{0j}(P_2)$-$C_{1j}(P_2)$ 를 비교하여 후자가 크면 $h_{1j}=1$로 하고 이외의 경우엔 $h_{2j}=1$로 한다. 이렇게 수정을 하면 h_1 과 h_2 는 g 와 양립하게 된다. 또한 위와 같은 결정 방법이 다른 어떤 결정 방법보다 일배체형을 유전자형과 양립시키면서 최소의 가중치를 들여 단편들을 분할하고 일배체형을 결정한다는 사실을 쉽게 알 수 있다.

일배체형의 결정 과정에 필요한 전체 가중치의 합을 계산하는 함수를 수정가중치함수라 하고 다음과 같이 정의한다.

$$FW(P)=\sum_{l=1}^{2}\sum_{f_i\in P_l} D(h_l,f_i) \quad (7)$$

여기서 단편들 f_i 는 수정하기 전의 분할된 단편들을 가리킨다. 달리 말하면 WMLF/GI 모델의 목적은 수정가중치함수의 값을 최소로 하는 행렬 M 의 분할을 찾는 것이다.

3.4 적응도 함수 설계

유전자 알고리즘에서는 세대내의 각 개체들의 적응도를 평가할 방법이 필요하다. 위에서 정의한 수정가중치함수를 적응도 함수로 사용해도 되지만 개체의 우수성이 높으면 적응도 함수도 높은 값을 갖도록 하는 것이 직관적 이해에 편리하므로 다음과 같이 약간의 수정을 가한 적응도 함수를 사용한다.

$$Fit((f_1,f_2,\ldots,f_m))=mn-FW(P) \quad (8)$$

적응도 함수는 $0<Fit((f_1,f_2,\ldots,f_m))\le mn$ 값을 갖는다. 여기서 적응도 함수를 mn 값으로 하는 개체 $(f_1,f_2,\ldots,f_m)$가 있으면 SNP 행렬은 타당하고 그 역도 성립함을 알 수 있다.

3.5 유전자 알고리즘 연산자

유전자 알고리즘엔 여러 가지 선택 연산자가 제시되어 있다. 그 중 수렴 속도와 정확성을 높이기 위해 토너멘트 선택 연산자와 세대내의 다양성을 보장하기 위해 교배 연산을 수행하는 룰렛휠 선택 연산자를 일정 비율로 함께 사용한다. 그리고 돌연변이는 한 점 돌연변이 연산을 사용한다. 이때 초기세대 생성시에 동일한 부분으로 분할된 단편들은 연산 후에도 동일한 값을 유지하도록 한다. 전체적인 유전자 알고리즘의 개요를 그림 3에 제시한다.

```
알고리즘 GAforHaplotypeAssembly
입력: SNP 단편 행렬 M, 가중치 행렬 W, 유전자형 g
      세대 크기 PS, 교배율 CR, 돌연변이율 MR,
      최대 세대 생성 수 GN
출력: 한 쌍의 일배체형 h1, h2

Begin
  임의의 초기 세대 P0 생성, k = 0;
  유전자형 g에 의해 초기 세대 수정
  while (k < GN) do
    세대 Pk 내의 각 개체들의 적응도 계산;
    토너멘트 선택 연산자를 이용하여 Pk 세대에서 (1-CR)×PS
    만큼의 개체들을 선택하여 Pk+1 세대에 편입;
    룰렛휠 선택 연산자와 교배연산자를 사용하여 Pk 세대에서
    CR×PS 만큼의 후손을 생성하여 Pk+1에 추가;
    새로 생성된 세대의 MR×PS 개체에 대해 돌연변이 수행
    k = k+1;
  end do
  return 적응도가 가장 큰 개체로 부터 결정한 한 쌍의 일배체형
end
```

그림 3 유전자 알고리즘 개요

4. 실험 결과 및 분석

제안한 알고리즘은 C 언어로 구현하고 32비트 시스템 (Pentium 4, 2.8 GHz 와 1GB RAM) 에서 실험하였다.

우리는 정확도 R_c를 모델과 알고리즘의 성능 평가치로 사용한다. 이 정확도는 다른 논문들[1,8,9,10]에서도 사용된 것으로 다른 모델이나 알고리즘과의 성능 비교를 위하여 그대로 사용한다. 정확도는 다음과 같이 정의한다.

$h^*=(h_1^*, h_2^*)$를 염색체에 대한 실제 일배체형이라 하고 $h=(h_1, h_2)$를 알고리즘에 의해 결정된 일배체형이라 하면 정확도 R_c는

$$R_c(h,h^*)=1-\frac{\min\{D(h_1,h_1^*)+D(h_2,h_2^*),D(h_1,h_2^*)+D(h_2,h_1^*)\}}{2n} \quad (9)$$

이다. 여기서 $D(h, h^*)$는 두 일배체형 간의 해밍 거리를 말한다.

4.1 임의 자료에 대한 실험

모델들의 성능을 비교하기 위해 길이가 n = 50인 50쌍의 종자(seed) 일배체형을 임의로 만들었다. 한 실험 개체의 SNP 행렬은 m = 50 개의 단편들로 구성했는데 이들은 한 쌍의 종자 일배체형을 임의로 복사하여 만들었다. 모든 SNP 단편들에는 손실률 R_m에 따라 임의로 손실을 발생시켰다. SNP 오류는 오류율 R_e에 따라 SNP 단편들의 임의 위치에 0은 1로 1은 0으로 수정하여 만들었다. 가중치 행렬 W는 정규분포를 따르도록 하였는데 정확한 SNP 위치에는 평균 μ = 0.9, 오류가 있는 위치에는 μ = 0.8이 되도록 하였다. 분산은 두 경우 모두 σ^2 = 0.05가 되도록 하였다. 설계한 유전자 알고리즘의 매개변수들은 PS = 400, CR = 0.8, MR = 0.01, GN = 150으로 하였다.

표 1은 WMLF 모델과 MLF/GI 모델 그리고 WMLF/GI 모델의 정확도를 오류율 0.05, 0.2, 0.3, 0.4 네 가지 경우에 대해 각각 손실률을 0.1~0.9하여 비교한 것이다. 표 안의 개별 값은 50개 실험개체의 정확도를 평균한 값이다. GI의 도입이 일배체형 조합의 정확도 측면에서 WMLF 모델에 미치는 영향이 큰 것을 알 수 있다. 또한 오류율과 손실률이 커지면 유전자형을 사용한 모델과 사용하지 않은 모델의 정확도 차가 더욱 커짐을 알 수 있다. 이러한 결과는 WMLF 모델에 유전자형 정보의 도입이 정확도의 개선에 효과가 있음을 보여준다. 본 실험에서는 WMLF/GI 모델과 MLF/GI 모델의 정확도가 큰 차이를 보이지는 않고 있다. 이는 신뢰치의 영향보다 유전자형의 영향이 상대적으로 크게 미친 결과로 여겨진다. 그러나 유전자형에 분포하는 이형의 구성비율에 따라 신뢰도의 영향을 고려할 때 정확도면에서 차이가 있을 것으로 여겨진다. 이는 추가적인 연구가 필요한 부분이다.

유전자 알고리즘의 효율성을 측정하기 위해 수렴세대수를 GN세대를 통해 최고의 정확성을 갖는 최초의 세대로 정의한다. 즉 수렴세대수는 GN세대 내에서 이후로 정확성의 개선이 더 이상 이루어지지 않는 세대를 가리킨다. 마찬가지로 50개의 실험 개체의 수렴세대수를 평균하였다. 표 2가 보여주듯 유전자형 정보의 도입이 MLF나 WMLF 모두 유전자 알고리즘 자체의 수렴 속도도 크게 향상시킴을 알 수 있다.

4.2 염색체 5q31에 대한 실험

Daly등[13]이 공개한 자료를 실험 자료로 사용하였다. 이 자료는 WMLF 문제를 휴리스틱 알고리즘인 동적 클러스터링 알고리즘으로 해결한 [8]에서도 사용하였다. 공개한 원자료는 부-모-자식 염색체 5q31의 103개 SNP위치에 대한 유전자형들로 구성되어 있다. 부모의 유전자형과 가계도 정보로부터 총 258쌍의 일배체형을 도출하였고 양 대립유전자를 정확히 결정할 수 없는 경우에는 손실로 처리하였다. 258쌍의 일배체형 중 손실율이 20%를 넘는 것들은 제거하고 나머지 147쌍의 일배체형을 실험 자료로 삼았다. 실험 개체를 생성하는 데는

표 1 임의 자료에 대한 세 모델의 정확도 비교

	R_e = 0.05			R_e = 0.2			R_e = 0.3			R_e = 0.4		
	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI
R_m=0.1	1.0	1.0	1.0	0.994	0.990	0.999	0.964	0.945	0.982	0.811	0.822	0.852
R_m=0.2	0.999	0.999	1.0	0.993	0.989	0.999	0.957	0.941	0.975	0.776	0.826	0.851
R_m=0.3	1.0	1.0	1.0	0.985	0.981	0.998	0.942	0.922	0.959	0.761	0.822	0.842
R_m=0.4	1.0	0.999	1.0	0.980	0.974	0.995	0.935	0.917	0.945	0.741	0.808	0.827
R_m=0.5	0.998	0.997	1.0	0.970	0.960	0.991	0.914	0.897	0.917	0.714	0.818	0.828
R_m=0.6	0.994	0.992	0.999	0.951	0.944	0.981	0.883	0.860	0.883	0.684	0.792	0.805
R_m=0.7	0.985	0.983	0.996	0.947	0.924	0.954	0.800	0.847	0.862	0.660	0.802	0.805
R_m=0.8	0.964	0.960	0.988	0.880	0.860	0.896	0.746	0.811	0.825	0.634	0.795	0.800
R_m=0.9	0.907	0.899	0.908	0.739	0.829	0.833	0.658	0.798	0.808	0.603	0.789	0.792

표 2 임의 자료에 대한 세 모델의 수렴세대수 비교

	R_e = 0.05			R_e = 0.2			R_e = 0.3			R_e = 0.4		
	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI
R_m=0.1	22.1	1.5	1.5	23.0	1.0	1.0	24.7	1.0	1.0	25.3	1.0	1.0
R_m=0.2	22.9	1.0	1.0	23.9	1.0	1.0	24.1	1.0	1.0	24.1	1.0	1.0
R_m=0.3	22.9	1.0	1.0	23.7	1.0	1.0	23.4	1.0	1.0	24.2	1.0	1.0
R_m=0.4	22.2	1.0	1.0	24.1	1.0	1.0	25.4	1.0	1.0	23.2	1.0	1.0
R_m=0.5	23.8	1.0	1.0	23.4	1.0	1.0	25.4	1.0	1.0	24.0	1.0	1.0
R_m=0.6	23.8	1.0	1.0	25.2	1.0	1.0	26.0	1.0	1.0	23.5	1.0	1.0
R_m=0.7	23.2	1.0	1.0	25.5	1.0	1.0	24.4	1.0	1.0	24.3	1.0	1.0
R_m=0.8	24.0	1.1	1.1	25.0	1.0	1.1	24.7	1.0	1.0	24.5	1.0	1.0
R_m=0.9	22.8	3.4	3.8	22.1	3.6	4.9	23.3	3.5	4.5	22.0	3.4	3.8

표 3 염색체 5q31에 대한 세 모델의 정확도 비교

	R_e = 0.05			R_e = 0.2			R_e = 0.3			R_e = 0.4		
	WMLF (wang)	MLF/GI (GA)	WMLF/GI (GA)	WMLF (wang)	MLF/GI (GA)	WMLF/GI (GA)	WMLF (wang)	MLF/GI (GA)	WMLF/GI (GA)	WMLF (wang)	MLF/GI (GA)	WMLF/GI (GA)
R_m=0.1	0.978	1.0	1.0	0.978	0.995	0.999	0.995	0.973	0.994	-	0.917	0.935
R_m=0.2	0.975	1.0	1.0	0.988	0.993	0.999	0.993	0.973	0.991	-	0.915	0.934
R_m=0.3	0.978	1.0	1.0	0.988	0.992	0.998	0.988	0.966	0.989	-	0.916	0.924
R_m=0.4	0.978	0.999	1.0	0.994	0.987	0.998	0.981	0.957	0.982	-	0.908	0.923
R_m=0.5	0.978	0.998	1.0	0.997	0.985	0.996	0.972	0.949	0.962	-	0.904	0.919
R_m=0.6	0.981	0.996	0.999	0.993	0.974	0.992	0.950	0.945	0.953	-	0.903	0.915
R_m=0.7	0.982	0.992	0.999	0.982	0.962	0.984	0.923	0.930	0.933	-	0.905	0.906
R_m=0.8	0.990	0.980	0.996	0.963	0.950	0.959	0.864	0.918	0.927	-	0.901	0.905
R_m=0.9	0.959	0.959	0.971	0.886	0.917	0.927	0.772	0.907	0.911	-	0.903	0.902

표 4 염색체 5q31에 대한 세 모델의 수렴세대수 비교

	R_e = 0.05			R_e = 0.2			R_e = 0.3			R_e = 0.4		
	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI	WMLF	MLF/GI	WMLF/GI
R_m=0.1	31.8	5.1	5.5	35.2	3.5	4.0	33.8	2.5	3.3	33.0	2.7	4.2
R_m=0.2	30.7	4.8	5.1	34.7	2.7	2.7	33.8	2.7	2.3	34.5	3.6	2.7
R_m=0.3	30.2	3.2	2.9	33.0	2.3	3.1	34.1	2.7	2.5	34.4	3.6	3.6
R_m=0.4	30.7	2.8	3.4	33.2	2.9	2.4	34.6	2.6	2.8	33.3	2.0	2.7
R_m=0.5	31.1	2.9	2.9	32.4	2.9	3.2	34.9	3.0	3.6	33.2	2.4	3.4
R_m=0.6	31.3	3.4	2.8	32.1	2.9	3.0	34.5	2.6	3.1	32.7	3.2	3.1
R_m=0.7	31.8	4.2	3.6	34.2	3.0	3.7	34.2	3.3	3.6	32.5	3.0	3.2
R_m=0.8	30.2	4.2	4.4	32.4	4.1	4.3	34.1	4.3	3.3	32.4	3.6	4.5
R_m=0.9	28.7	6.4	6.2	32.2	5.8	7.1	31.9	5.4	6.3	31.7	5.8	6.5

m = 100개의 SNP 단편들로 SNP 행렬을 구성한 것을 제외하면 임의 자료에 대한 것과 동일한 매개변수들을 사용하였다. 이는 [13]에서 사용한 것과 동일하다.

생성한 자료에 대한 세 모델의 정확도를 표 3에서 비교 하였다. 결과치의 첫 번째 열은 [8]의 결과이다. 마찬가지로 표 안의 값들은 147개의 실험개체들을 각각의 매개변수들에 따라 실험한 후의 정확도를 평균한 값이다. 임의 자료의 실험에서와 마찬가지로 WMLF 모델에 유전자형 정보의 도입이 손실률과 오류율이 높을수록 정확도를 개선하는데 더 효과적임을 알 수 있다. 그리고 WMLF/GI 모델이 작지만 MLF/GI 모델보다 성능이 나음도 알 수 있다.

표 4는 실제 자료에 대한 세 모델의 수렴속도를 비교한 것이다. 유전자형 정보의 도입이 가중치 행렬의 사용 여부와 상관없이 유전자 알고리즘의 수렴 속도를 크게 향상시킴을 보여준다.

5. 결 론

일배체형 조합 문제는 전산생물학분야에서 중요한 문제 중 하나이다. 본 논문에서는 WMLF모델을 개선한 WMLF/GI 모델을 제시하고 이 문제가 NP-hard임을 증명하였다. 그리고 문제의 정확성을 높이고 효율적으로 해결하기 위한 유전자 알고리즘을 제시하였다. 알고리즘은 유전자형 정보를 일배체형의 초기 세대 형성과 선정에 사용함으로써 알고리즘의 정확성과 수렴속도를 증가시켰다. 임의 자료와 실제 인간 염색체에 대한 실험 결과는 알고리즘이 이러한 목적을 잘 달성하고 있으며 WMLF 모델 보다 WMLF/GI 모델이 특히 SNP 단편들에 손실과 오류가 많은 경우 훨씬 높은 정확도를 가짐을 보여주었다. 현재의 실험 결과는 WMLF/GI 모델과 MLF/GI 모델의 정확도면에서의 차이가 크지 않음을 보였는데 이는 유전자형 정보의 사용 여부가 판독상의 신뢰치의 사용보다 정확도와 수렴속도 면에서 훨씬 큰 영향을 미침을 알 수 있다. 그러나 유전자형에 분포하는 이형의 구성비율에 따라 영향은 달라질 것으로 보이며 따라서 모델간의 정확도면에서도 차이가 있을 것이다. 이는 두 모델간의 성능 비교를 위한 중요한 요소이며 WMLF/GI 모델에 대한 새로운 접근 방법을 개발하고 이들의 결과를 바탕으로 두 모델에 대한 종합적인 성능 비교가 필요하다.

참 고 문 헌

[1] R. S. Wang, L. Y. Wu, Z. P. Li, and X. S. Zhang, "Haplotype reconstruction from SNP fragments by minimum error correction," *Bioinformatics*, Vol. 21, No. 10, pp. 2456-2462, 2005.

[2] J. D Terwilliger and K. M Weiss, "Linkage disequilibrium mapping of complex disease: fantasy or reality?," *Current Opinion in Biotechnology*, Vol. 9, No. 6, pp. 578-594, 1998.

[3] J. C. Stephens, et al, "Haplotype variation and

linkage disequilibrium in 313 human genes," *Science*, vol. 293, pp. 489-493, 2001.
[4] X. S. Zhang, R. S. Wang, L. Y. Wu, and L. Chen, "Models and Algorithms for Haplotyping Problem," *Current Bioinformatics*, Vol. 1, pp. 105-114, 2006.
[5] H. J. Greenberg, W. E. Hart, and G. Lancia, "Opportunities for Combinatorial Optimization in Computational Biology," *INFORMS Journal on Computing*, Vol. 16, No. 3, pp. 211-231, 2004.
[6] R. Cilibrasi, L. V. Iersel, S. Kelk, and J. Tromp, "On the complexity of Several Haplotyping Problem," *5th Workshop on Algorithms in Bioinformatics*(WABI), LNBI 3692, pp. 128-139, 2005.
[7] R. Rizzi, V. Bafna, S. Istrail, and G. Lancia, "Practical Algorithms and Fixed-Parameter Tractability for the Single Individual SNP Haplotyping Problem," *2nd Workshop on Algorithms in Bioinformatics(WABI)*, LNCS 2452, pp. 29-43, 2002.
[8] Y. Y. Zhao, L. Y. Wu, J. H. Zhang, R. S. Wang, and X. S. Zhang, "Haplotype assembly from aligned weighted SNP fragments," *Computational Biology and Chemistry*, Vol. 29, pp. 281-287, 2005.
[9] Y. Wang, E. Feng, R. Wang, and D. Zhang, "The haplotype assembly model with genotype information and iterative local-exhaustive search algorithm," *Computational Biology and Chemistry*, Vol. 31, pp. 288-293, 2007.
[10] X. S. Zhang, R. S. Wang, L. Y. Wu, and W. Zhang, "Minimum Conflict Individual Haplotyping from SNP Fragments and Related Genotype," *Evolutionary Bioinformatics Online*, Vol. 2, pp. 271-280, 2006.
[11] S. H. Kang, I. S. Jeong, M. H. Choi, and H. S. Lim, "Haplotype Assembly from Weighted SNP Fragments and Related Genotype Information," *Frontiers in Algorithmics Workshop(FAW) 2008*, LNCS 5059, pp. 45-54, 2008.
[12] D. E. Goldberg, *Genetic Algorithms in serarch, Optimization and Machine Learning*, Addison-Wesley, 1989.
[13] 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 29*, pp. 229-232, 2001.

강 승 호
1994년 전남대학교 전산학과 학사. 2003년 전남대학교 전산학과 석사. 2003년~현재 전남대학교 전산학과 박사과정. 관심분야는 생물정보학, 알고리즘, 인공지능 등임

정 인 선
2001년 여수대학교 전산학과 학사. 2006년 전남대학교 전산학과 석사. 2006년~현재 전남대학교 전산학과 박사과정. 관심분야는 생물정보학, 알고리즘, 인공지능 등임

최 문 호
1993년 전남대학교 전산학과 학사. 1995년 전남대학교 전산학과 석사. 2001년~현재 전남대학교 전산학과 박사과정. 관심분야는 알고리즘, 생물정보학 등임

임 형 석
1983년 서울대학교 컴퓨터공학과 학사
1985년 한국과학기술원 전산학과 석사
1993년 한국과학기술원 전산학과 박사
1996년~1997년 미국 Purdue대학 방문교수. 1987년~현재 전남대학교 전자컴퓨터공학부 교수. 관심분야는 알고리즘, 그래프이론, 생물정보학 등임