DOI: 10.3745/KIPSTB.2009.16-B.2.157

일배체형 재조합을 위한 MCIH 모델과 WMLF/GI 모델의 정확도 비교

정 인 선[†] · 강 승 호[†] · 임 형 석[††]

요 약

일배체형 조합 문제를 해결하기 위해 제시된 MLF(Minimum Letter Flips) 모델이나 WMLF(Weighted Minimum Letter Flips) 모델은 유전자형 정보를 도입함으로써 오류와 손실이 많을 때에도 높은 정확도를 얻을 수 있다. 그리고 MLF 모델에 비해 가중치 버전인 WMLF모델의 정확도가 높다는 사실도 밝혀졌다. 본 논문에서는 유전자형 정보상의 동형(homozygous)의 분포 비율과 유전자 서열판독기계의 성능에 따른 신뢰도의 차이를 매개변수로 하여 두 모델을 구체적으로 비교, 분석한다. 두 모델의 성능 비교를 위해 신경망과 유전자 알고리즘을 사용한다. 실험결과 동형의 비율이 크고 판독기계의 성능이 좋으면 특히 손실율과 오류율이 높은 경우에 WMLF/GI 모델의 정확도가 더 우수함을 보인다.

키워드 : 일배체형 조합문제, 유전자형, SNP, 신경망

The Correctness Comparison of MCIH Model and WMLF/GI Model for the Individual Haplotyping Reconstruction

In-Seon Jeong[†] · Seung-Ho Kang[†] · Hyeong-Seok Lim[††]

ABSTRACT

Minimum Letter Flips(MLF) and Weighted Minimum Letter Flips(WMLF) can perform the haplotype reconstruction more accurately from SNP fragments when they have many errors and gaps by introducing the related genotype information. And it is known that WMLF is more accurate in haplotype reconstruction than those based on the MLF. In the paper, we analyze two models under the conditions that the different rates of homozygous site in the genotype information and the different confidence levels according to the sequencing quality. We compare the performance of the two models using neural network and genetic algorithm. If the rate of homozygous site is high and sequencing quality is good, the results of experiments indicate that WMLF/GI has higher accuracy of haplotype reconstruction than that of the MCIH especially when the error rate and gap rate of SNP fragments are high.

Keywords : Haplotype Assembly Problem, Genotype, SNP, Neural Network

1. 서 론

사람의 유전체(genome) 서열이 전부 밝혀짐에 따라 유전적 차이에 대한 연구가 유전학에서 중요한 주제가 되었다[10]. 모든 사람은 개인 간에 약 99.9% 동일한 유전적 염기서열을 지니고 있고 약 0.1%의 염기서열만이 개인 간의 차이를 보인다. 단지 0.1%의 염기서열 차이가 개인간 유전적 차이의 원인으로 추정되고 있다[8]. 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 DNA 염기서열에서 하나의 염기서열 차이를 보이는 유전적 변화 또는 변이를 말한다. 변이의 범위는 제한되어 있고 각 변이를 대립유전자(allele)라 한다. SNP은 일반적으로 대립유전자의 빈도에 의해 원형(wild type)과 돌연변이형(mutant type)으로 구분한다. 그리고 특정 염색체의 SNP 서열을 일배체형이라 한다. 인간과 같은 이배체(diploid) 생물은 유전체가 한 쌍의 염색체로 구성되어 있기 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 동형이라 하고, 서로 다르면 이형(heterozygous)

※ 이 논문은 2007년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2007-313-D00644).
† 준 회 원 : 전남대학교 전산학과 박사과정
†† 정 회 원 : 전남대학교 전자컴퓨터공학부 교수
논문접수 : 2009년 1월 5일
수 정 일 : 1차 2009년 2월 24일
심사완료 : 2009년 2월 25일

	SNP 1	SNP 2	SNP 3
Chrom. *c*, paternal	g c g g **T** a a g ··· a c g **G** c t a ··· g g a t **C** a g t		
Chrom, *c*, maternal	g c g g **A** a a g ··· a c g **G** c t a ··· g g a t **G** a g t		
일배체형 1	T	G	C
일배체형 2	A	G	G
유전자형	T/A	G/G	C/G

(그림 1) 한 쌍의 일배체형과 유전자형 예

이라 한다. (그림 1)은 3개의 SNP 위치를 갖는 염색체의 예이다. 일배체형들은 {TGC}와 {AGG}이며, 이형은 SNP 위치 1과 3, 동형은 SNP 위치 2에서 나타난다.

일배체형의 차이는 개별 개체들의 발현 형질의 차이와 직접적인 관련이 있는 것으로 밝혀져 있다. 특히 유전과 관련이 있는 질병연구에서 중요한 의미를 갖고 있다. 이들의 차이를 유형화 하고 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 일배체형 유형의 관계를 규명하게 되면 맞춤의학의 실현 또한 가능할 것으로 판단된다. 이처럼 일배체형은 생물학, 의학, 약학 등 여러 분야에서 중요한 의미를 갖는 유전마커이다[7].

개별 SNP이나 유전자형을 생물학적 실험을 통해 얻어내기는 상대적으로 쉬운 작업에 속하지만 일배체형을 바로 얻어내는 것은 기술적 제약이 많으며 비싼 비용을 지불해야 한다. 또한 현재의 SNP 판독 기술은 SNP의 위치와 대립유전자형을 알아낼 수 있을 뿐 대립유전자가 두 염색체 중 어떤 염색체의 것인지를 판별하기에 한계가 있다. 따라서 일배체형을 결정하는 문제가 유전자형을 결정하는 문제보다 훨씬 어렵다. 이러한 어려움을 극복하기 위해 전산학적 관점에서 두 부류의 문제가 정의 되었다. 일배체형 추론문제(haplotype inference problem)는 특정 집단의 유전자형 정보로부터 일배체형 집합을 추론하는 문제이다. 다른 하나는 일배체형 조합문제(haplotype assembly problem)로 개인의 판독된 염기 서열들로부터 하나의 일배체형 쌍을 조합해내는 문제이다. 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 조건들을 가정하고 있다[11]. 이중 MLF 모델과 이 모델의 가중치 버전인 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 염기를 판독하는 기계가 자신이 판독한 개별 염기에 대해 신뢰도를 부여한다는 사실에 기반하고 있다. 이 가중치는 SNP 판독 기계가 판독된 염기에 대해 가지는 정확성을 나타낸다. 두 모델이 제시한 문제들은 단편들에 손실이 없는 경우에도 NP-hard임이 증명되었으며[1,13], 1개의 손실이 있는 경우엔 MLF 모델은 APX-hard임[1]이 밝혀졌다. APX-hard란 좋은 근사 알고리즘의 존재가 알려지지 않은 경우를 말한다. 그리고 Zhao 등[13]은 WMLF 모델이 MLF 모델 보다 일배체형을 조합하는데 높은 정확성을 가짐을 보여주었다. 그러나 두 모델은 SNP 판독상의 손실률과 오류율이 낮은 경우에 효과적이다. MLF 모델의 정확성을 향상시키기 위해 상대적으로 얻기 쉬운 유전자형 정보의 도입을 모델화 한 MCIH 모델[12]이 제시되었다. MCIH 모델은 MLF/GI 모델[10]로 불린다. 이후 유전자형 정보를 MLF 모델에 도입한 여러 방법들이 제시되었다[9,10,12]. [14]에서는 WMLF 모델에 유전자형을 도입한 WMLF/GI 모델을 제시하고 일배체형 결정의 정확도와 수렴속도를 향상시켰다. 그러나 유전자형을 도입한 두 모델 WMLF/GI와 MCIH 사이의 유전자형 정보에 따른 다양한 비교는 되어있지 않고 유전자형에 분포하는 동형의 비율과 서열판독기계의 성능에 따른 두 모델간의 성능 비교를 후속 연구로 제시하고 있다. 본 논문은 이러한 사실을 바탕으로 [14]에서 제시한 두 가지 조건을 사용하여 두 모델의 성능상의 차이를 종합적으로 보인다. 특히 이 두 가지 조건이 어떻게 두 모델간의 일배체형 조합의 정확도에 기여하는지를 보인다. 이를 위해 새로이 신경망을 설계하고 기존에 제시된 유전자 알고리즘 두 가지를 이용하여 두 모델의 성능상의 차이를 보다 구체적으로 비교, 분석한다.

논문의 구성은 다음과 같다. 2장에서는 문제에 대한 정의를 보이고, 3장에서는 제시된 문제를 해결하기 위한 알고리즘을 설계한다. 마지막으로 4장과 5장에서 실험결과에 대해 분석하고 결론을 맺는다.

2. 문제 정의

실험을 통해 한 쌍의 염색체로부터 길이가 n인 m개의 SNP 단편들을 얻었다고 하자. 각 SNP는 원형이거나 돌연변이형 혹은 손실일 수 있으며, 각각 1, -1 그리고 0으로 표기한다. 이러한 단편들은 {1, -1, 0}로 구성된 $m \times n$ 행렬 M로 표현되는데, 이를 SNP 행렬이라 부른다. 행렬의 각 행은 SNP 단편 f_i에 해당하고 각 열은 단편들의 SNP 위

치에 해당한다. SNP 판독기계는 이러한 단편들의 SNP 값에 신뢰도를 부여하는데 이는 SNP 값이 올바로 판독 되었는지에 따라 0과 1사이의 확률 값으로 나타낸다. SNP 값에 대한 신뢰도는 $m \times n$ 의 가중치 행렬 W 로 표현되고 행렬의 원소 w_{ij} 는 행렬 M 의 원소인 SNP 위치 f_{ij} 의 값에 대한 신뢰도를 나타낸다. f_{ij} 가 손실이면 신뢰도를 0으로 한다. 서로 다른 값을 가진 SNP 위치는 가중치가 낮은 값을 다른 값으로 바꾸면 적은 비용으로 일치시킬 수 있으므로 두 단편 f_i 와 f_j 의 SNP 위치 사이의 거리는 그들의 가중치를 사용하여 다음과 같이 정의한다.

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

만약 한쪽이 SNP 단편이 아니고 일배체형인 경우엔 SNP 단편의 가중치를 사용한다. 즉, $d(f_{ik}, h_{jk}) = w_{ik}$ 이다. SNP 단편 $f_i = (f_{i1}, \ldots, f_{in})$ 과 $f_j = (f_{j1}, \ldots, f_{jn})$ 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_i 과 f_j 가 다른 염색체에서 복제되었거나 SNP 판독에 오류가 있었음을 의미하고 이런 경우를 충돌한다(conflict)라고 한다. 그렇지 않으면 모든 SNP 단편들이 서로소인 두 집합으로 분리되고 집합내의 모든 단편들 간에 충돌이 없음을 의미하는데 이때 SNP 행렬이 타당하다(feasible)라고 한다. 일배체형 h_i 와 SNP 단편 f_j 사이의 거리는 아래와 같이 정의한다.

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

그리고 (그림 1)처럼 유전자형 $g = (g_1, g_2, \ldots, g_n)$ 에 대해 i 번째 SNP 위치가 모두 원형의 대립유전자를 가지면 2를 g_i 에 부여하고 돌연변이형의 대립유전자를 가지면 -2를 부여한다. 만약 SNP 위치가 이형이면 0을 부여한다. 그리고 한 쌍의 일배체형 h_1 과 h_2 가 각 SNP 위치에 대해서 아래의 조건을 만족하면 이 일배체형 쌍은 유전자형과 양립한다(compatible)라고 한다.

$$\begin{aligned} &\text{if } g_i = 2, \quad h_{1i} = h_{2i} = 1 \\ &\text{if } g_i = -2, \quad h_{1i} = h_{2i} = -1 \\ &\text{if } g_i = 0, \quad h_{1i} = -h_{2i} = -1 \text{ or } h_{1i} = -h_{2i} = 1 \end{aligned} \quad (4)$$

MCIH 문제는 유전자형 정보를 갖는 SNP 행렬이 주어졌을 때 행렬의 원소 값들을 최소 개수로 변경하여 두 집합이 타당하도록 분할하고 유전자형과 양립하는 두 일배체형을 결정한다. WMLF/GI 문제는 MCIH 문제에 서열판독기계의 성능에 따른 신뢰도를 추가하여 두 일배체형을 결정한다. 이에 대한 WMLF/GI 문제는 다음과 같이 정의한다[5].

정의 1. WMLF/GI 문제 *SNP 행렬 M 과 가중치 행렬 W 그리고 유전자형 g 가 주어지면, "가중치의 합이 최소"이면서 변경 후의 SNP 행렬이 "타당"하고 유전자형과 "양립"하도록 SNP 행렬의 원소 값들을 1에서 -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.*

3. WMLF/GI문제를 해결하기 위한 알고리즘 설계

3.1 신경망을 이용한 일배체형 조합 문제

WMLF/GI 문제는 분류 문제와 유사하다. 즉 SNP 단편들이 주어졌을 때 각 SNP 단편을 집합 내의 단편들끼리 상호 충돌이 최소가 되도록 두 집합으로 분리하고 이 두 집합으로부터 한 쌍의 일배체형을 결정한다. 신경망은 잡음이 많은 데이터와 훈련 받지 않은 데이터에 대해 분류 능력이 우수하다고 알려져 있으며 생물정보학 분야를 포함한 여러 분야에서 성공적으로 사용되고 있다.

신경망은 (그림 2)처럼 3개의 계층(layer)으로 구성된다. 입력계층에서 각 노드는 길이가 n 인 m 개의 SNP 단편들로써 {1, -1, 0} 값으로 구성된 n-차원의 벡터이다. 은닉계층은 두 개의 노드를 가지며 각 노드는 한 쌍의 일배체형에 대응되는 단편들의 두 부분집합을 나타낸다. 출력계층은 하나의 노드로 유전자형을 나타낸다. 신경망의 중요한 특징은 다음의 목적을 성취하기 위해 설계되었다.

목적 1. 최소화

$$\sum_{l=1}^{2} \sum_{f_i \in C_l} D(h_l, f_i) \quad (5)$$

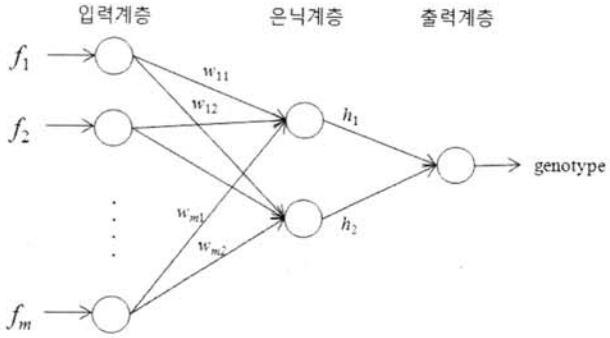
여기서 $i = 1, 2, \ldots, m$ 이다. $f = (f_1, f_2, \ldots, f_m)$ 의 분할 $P = (C_1, C_2)$ 와 한 쌍의 일배체형 (h_1, h_2) 에 대해 C_l 에 속한 f_i 와 h_l 사이의 전체 가중치의 합을 최소화 한다.

목적 2. 만족

$$z_i = g_i, \quad i = 1, \ldots, n \quad (6)$$

또는 최소화

$$\sum_{i=1}^{n} (z_i - g_i), \quad i = 1, \ldots, n \quad (7)$$



(그림 2) 3계층의 신경망

실험을 통해 얻은 유전자형 $Z=(z_1,z_2,...,z_n)$와 실제 유전자형 $g=(g_1,g_2,...,g_n)$이 동일하거나 그 차를 최소화 한다.

3.1.1 연결 가중치의 초기화

입력계층에서 은닉계층의 연결 가중치 $W'=(w'_{ij})$는 0 또는 1의 임의의 값을 갖는다. 여기서 $i=1,2$이고 $j=1,2,...,m$이다.

3.1.2 일배체형 생성을 위한 전방향 처리

전방향 처리는 은닉계층과 출력계층에 있는 각 노드의 입력값과 출력값들의 계산에 의해 한쌍의 일배체형과 유전자형을 생성하며 처리과정은 다음과 같다.

1) 은닉계층의 입력값은 h_1, h_2 각각에 대해 $X_1=(x_{11},x_{12},...,x_{1n})$와 $X_2=(x_{21},x_{22},...,x_{2n})$이다. 각 노드의 입력을 계산하기 위해서 이 노드에 연결된 입력을 해당 가중치와 곱한 후에 합한다.

$$x_{lk}=\sum_{i=1}^{m} f_{ik}w'_{il},\ \ l=1,2,\ k=1,...,n. \tag{8}$$

2) 은닉계층의 출력값은 한 쌍의 일배체형 $h_1=(h_{11},h_{12},...,h_{1n})$과 $h_2=(h_{21},h_{22},...,h_{2n})$이다. 은닉계층은 입력값을 받아서 활성화 함수를 적용한다. 활성화 함수는 입력값의 범위를 -1과 1 사이의 값으로 대응시키며 다음과 같다.

$$h_{lk}=F(x_{lk})=\frac{2}{1+e^{-\lambda x}}-1. \tag{9}$$

3) 출력계층은 h_1과 h_2을 입력으로 받아 아래 식에 의해 유전자형 $Z=(z_1,z_2,...,z_n)$을 출력한다.

$$z_k=h_{1k}+h_{2k}. \tag{10}$$

3.1.3 오류 역전파 알고리즘을 통한 연결 가중치 갱신

모든 SNP 단편들을 전방향 처리를 통해 얻어진 h_1과 h_2 사이의 거리에 의해 두 집합 $P=(C_1,C_2)$으로 분류한다. 만약 $D(h_1,f_i)<D(h_2,f_i)$ 라면, f_i는 C_1에 분류되며 그렇지 않으면 C_2로 분류된다. 일배체형 h_l과 SNP 단편 f_i 사이의 거리는 다음과 같이 정의한다.

$$D(h_l,f_i)=\sum_{k=1}^{n} d(s(h_{lk}),f_{ik}), \tag{11}$$

$$s(x)=\begin{cases}+1, & x\ge 0\\ -1, & x<0.\end{cases}$$

여기서 $l=1,2$이고 $k=1,...,n$이다. 거리 d는 식 (1)에 의해 정의되었으며, 한쪽이 SNP 단편이 아니고 일배체형이므로 둘 사이의 거리는 SNP 단편의 가중치를 사용한다. 즉 $d(s(h_{lk}),f_{ik})=w_{ik}$이다.

식 (12)는 단편들의 부분집합 C_l에 속하는 SNP 단편들과 일배체형 h_l 사이의 오류이며, 식 (13)은 실험을 통해 얻은 유전자형 $Z=(z_1,z_2,...,z_n)$와 실제 유전자형 $g=(g_1,g_2,...,g_n)$ 사이의 오류이다. 이들 오류를 최소화하기 위해 은닉계층과 출력계층 사이의 연결가중치를 갱신한다.

$$Err\_h_l=\sum_{f_i\in C_l}\sum_{k=1}^{n}(h_{lk}-f_{ik})^2\left|f_{ik}\right|,\quad l=1,2,\ i=\ 1,2,\ ...\ m \tag{12}$$

$$Err\_g=\sum_{k=1}^{n}(z_k-g_k)^2. \tag{13}$$

연결가중치 $W'(t)$의 갱신은 다음 식에 의해 계산한다.

$$w'_1(t+1)=w'_1(t)-\rho\left(L_1\frac{\partial Err\_h_1}{\partial w_1}+L_2\frac{\partial Err\_g}{\partial w_1}\right), \tag{14}$$

$$w'_2(t+1)=w'_2(t)-\rho\left(L_1\frac{\partial Err\_h_2}{\partial w_2}+L_2\frac{\partial Err\_g}{\partial w_2}\right), \tag{15}$$

연결가중치 $w'_l=(w_{1l},w_{2l},...,w_{ml})^T$에 대한 Err_h_l

와 Err_g 는

$$\frac{\partial Err\_h_l}{\partial w_l} = \begin{cases} \sum_{k=1}^{n} \lambda / 2[h_{lk} - f_{ik}] \cdot [1 - h_{lk}^{\ 2}] f_{ik}, & if\ f_i \in C_l \\ 0 & , if\ f_i \notin C_l \end{cases} \quad (16)$$

$$\frac{\partial Err\_g}{\partial w_l} = \begin{cases} \sum_{k=1}^{n} \lambda / 2[z_{lk} - g_{lk}] \cdot [1 - h_{lk}^{\ 2}] f_{ik}, & if\ f_i \in C_l \\ 0 & , if\ f_i \notin C_l \end{cases} \quad (17)$$

이다. 여기서, $l = 1, 2,\ i = 1, 2, \ldots m$, ρ 는 학습률이며 나머지는 신경망을 학습하기 위한 매개변수들이다.

3.2 유전자 알고리즘

유전자 서열판독기계의 성능에 따른 신뢰도와 유전자형 내의 동형 분포율에 의해 한 쌍의 일배체형을 결정하는 문제를 실험하기 위해 이전에 제안된 유전자 알고리즘을 간략하게 제시한다[5].

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 h_1, h_2

Begin
임의의 초기 세대 P_0 생성, $k = 0$;
유전자형 g에 의해 초기 세대 수정
while $(k < GN)$ *do*
세대 P_k 내의 각 개체들의 적응도 계산 ;
토너멘트 선택 연산자를 이용하여 P_k 세대에서
만큼의 개체들을 선택하여 P_{k+1} 세대에 편입;
룰렛휠 선택 연산자와 교배연산자를 사용하여
$CR \times PS$ 만큼의 후손을 생성하여 P_{k+1}에 추가;
새로 생성된 세대의 $MR \times PS$ 개체에 대해 돌연
$k = k + 1$;
end do
return 적응도가 가장 큰 개체로 부터 결정한 한
end

(그림 3) 유전자 알고리즘 개요

4. 실험 결과 및 분석

일배체형 조합을 위한 WMLF/GI 모델의 성능을 평가하기 위해 실제 데이터와 임의 데이터를 사용하였다. 제안한 알고리즘은 C 언어로 구현하고 32비트 시스템(Pentium 4, 3.20 GHz 와 2GB RAM)에서 실험하였다.

정확도 R_r (Reconstruction rate)를 모델과 알고리즘의 성능 평가치로 사용한다. 이 정확도는 다른 논문들[9,10,12,13]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 $\hat{h} = (\hat{h}_1, \hat{h}_2)$ 를 알고리즘에 의해 결정된 일배체형이라 하면 정확도 R_r 는

$$R_r(h, \hat{h}) = 1 - \frac{\min\{r_{11} + r_{22},\ r_{12} + r_{21})}{2n} \quad (18)$$

이다. 여기서 $r_{ij} = D(h_i, \hat{h}_j),\ i = j = 1,\ 2$ 로써 두 일배체형 간의 해밍 거리를 말한다.

4.1 임의 자료에 대한 실험

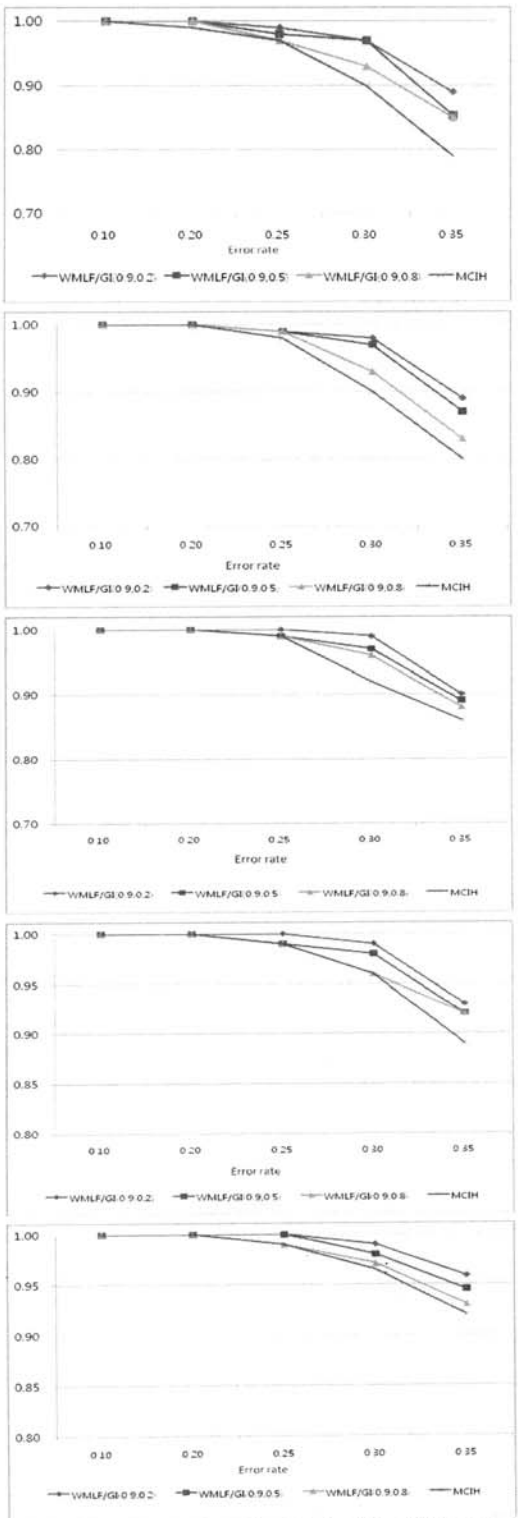
모델들의 성능을 비교하기 위해 길이가 n =50인 20쌍의 종자(seed) 일배체형을 임의로 만들었다. 한 실험 개체의 SNP 행렬은 m =50개의 단편들로 구성했는데 이들은 한 쌍의 종자 일배체형을 임의로 복사하여 만들었다. 모든 SNP 단편들에는 손실률(R_m =0.3)에 의해 임의로 손실을 발생시켰다. SNP 오류는 오류율(R_e =0.1, 0.2, 0.25, 0.3, 0.35)에 따라 SNP 단편들의 임의 위치에 -1은 1로, 1은 -1로 수정하여 만들었다. 설계한 신경망 알고리즘의 매개변수들은 ρ = 0.03, λ = 0.1, L_1 =0.2, L_2 =0.8로 하였다.

앞에서 언급 한데로 일배체형 조합문제에 대해 오류율과 손실률이 높을수록 WMLF 모델이 MLF 모델 보다 정확성이 높다. 유전형 정보를 추가한 MCIH 모델과 WMLF/GI 모델의 성능을 비교하기 위해 필요한 가정과 매개변수를 추가하였다. WMLF 또는 WMLF/GI 모델은 유전자의 서열을 판독하는 기계가 판독에 대해 신뢰도를 부여한다는 사실을 전제하고 있다. 이때 판독 기계의 성능이 우수하다면 오류가 있는 판독은 신뢰도를 낮게 부여하고, 오류가 없는 경우에 대해서는 높은 신뢰도를 부여할 가능성이 크다. 따라서 기계의 성능에 따라 판독에 대한 신뢰도의 차이는 현실적인 가정을 해치지 않는다. 한편 유전자형 정보의 도입은 유전자형이 가지고 있는 동형의 분포율에 따라 일배체형 조합의 정확성에 매우 큰 영향을 미친다.

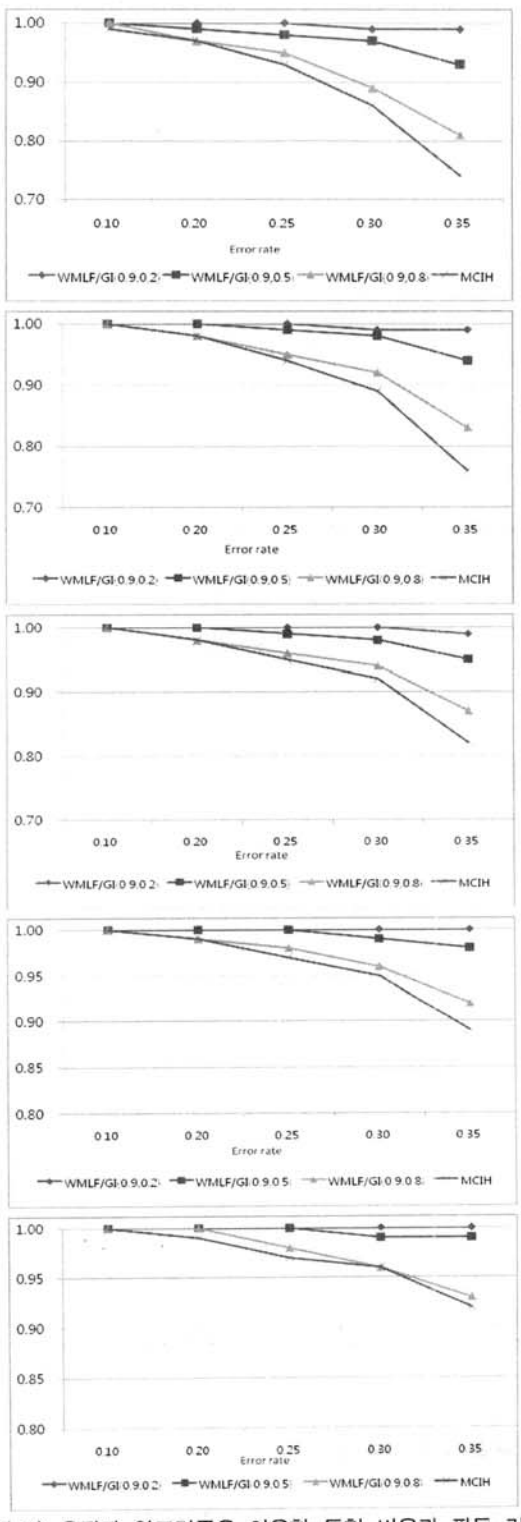
본 논문에서는 판독에 있어 신뢰도의 차이로 대변되는 판독기계의 성능이라는 가정과 유전자형 내의 동형의 분포율이라는 매개변수를 이용하여 MCIH 모델과 WMLF/GI 모델의 성능을 비교하였다. WMLF/GI 모델에서 판독기계의 성능에 따라 신뢰도(0.9, 0.2), 신뢰도(0.9, 0.5), 신뢰도(0.9, 0.8) 세가지로 실험하였다. 첫 번째 숫자는 판독이 정확한 경우

에 부여한 신뢰도의 평균이며, 두 번째 숫자는 그 반대의 신뢰도의 평균이다.

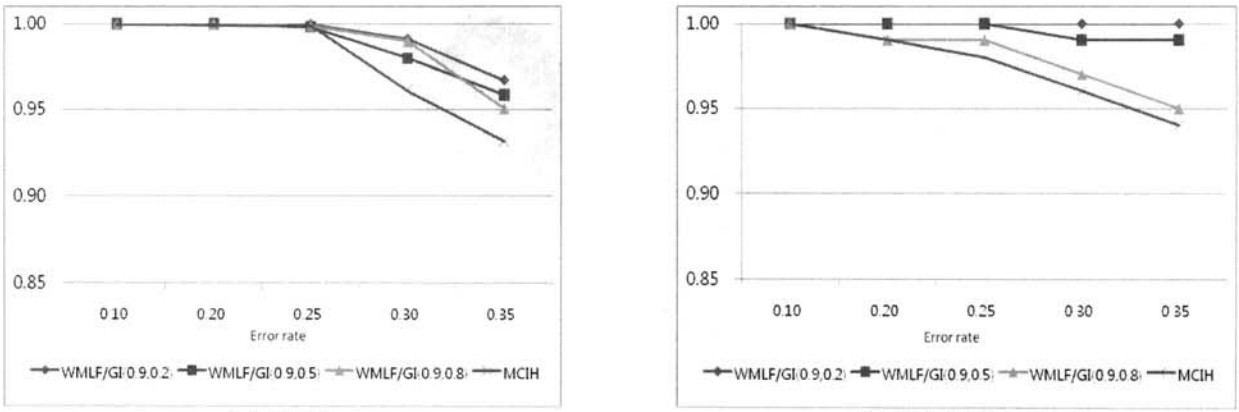
(그림 4)와 (그림 5)는 신경망과 유전자 알고리즘[5]을 이용하여 얻은 두 모델의 정확도로써 실험 개체들을 각각의 매개변수들에 따라 실험한 후에 정확도를 평균한 값이다. 신뢰도를 사용한 WMLF/GI 모델이 MCIH 모델 보다 높은 정확도를 가짐을 알 수 있다. 신뢰도(0.9,0.2)이고 오류율이 0.3이내인 경우 신경망은 97%이상 그리고 유전자 알고리즘



(그림 4) 신경망을 이용한 동형 비율과 판독 기계의 성능에 따른 정확도 비교
(동형 비율: 0%, 10%, 30%, 60%, 70%)



(그림 5) 유전자 알고리즘을 이용한 동형 비율과 판독 기계의 성능에 따른 정확도 비교
(동형 비율: 0%, 10%, 30%, 60%, 70%)



(a) 신경망 (b) 유전자 알고리즘

(그림 6) MCIH모델과 WMLF/GI모델의 정확도 비교

은 99%이상의 정확도를 보였다. 오류율이 0.35인 경우 신경망은 신뢰도를 사용하지 않은 MCIH 보다 5~10% 정도의 정확도를 향상 시켰으며, 유전자 알고리즘은 동형의 비율이 0%일 때 MCIH 보다 25%이상의 높은 정확도를 보였다. 신뢰도의 차이가 큰 경우, 즉 서열판독기계의 정확성이 높을 때 일배체형 결정의 정확도가 높음을 알 수 있다.

그리고 유전자형 정보가 정확도에 미치는 영향은 동형의 비율에 크게 좌우되며, 동형의 비율이 낮으면 상대적으로 서열판독기계의 성능이 일배체형 결정에 영향이 커진다는 사실을 알 수 있다. 특히 유전자 알고리즘을 이용한 실험에서는 동형의 비율이 낮을수록 상대적으로 기계 성능의 영향력이 커지고 오류율이 높을수록 일배체형 결정의 정확도의 차가 커짐을 알 수 있다. 하지만 신경망의 경우는 유전자형 내의 동형에 보다 민감하게 반응하는 반면 판독기계의 성능에는 크게 반응하지는 않는 것으로 나타났다.

4.2 염색체 5q31에 대한 실험

Daly등[2]이 공개한 자료를 실험 자료로 사용하였다. 공개한 원자료는 부-모-자식 염색체 5q31의 103개 SNP 위치에 대한 유전자형들로 구성되어 있다. 부모의 유전자형과 가계도 정보로부터 총 258쌍의 일배체형을 도출하였고 양대립유전자를 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 m=50개의 SNP 단편들로 SNP 행렬을 구성하였다.

생성한 자료에 대한 MCIH 모델과 WMLF/GI 모델의 정확도를 (그림 6)에서 비교 하였다. 실제 염색체 5q31에 대한 일배체형 258쌍의 동형 비율은 평균 76.5% 정도였다. SNP 판독 기계 성능에 따른 신뢰도 및 매개변수들은 임의 자료와 동일하게 실험하였다. 실제 데이터에서도 임의 데이터와 마찬가지로 WMLF/GI 모델이 전체적인 우위를 보였다. 오류율이 0.2 이내로 낮은 경우 두 모델의 정확도는 비슷하나, 오류율이 큰 경우 SNP 단편들에 신뢰도를 부여하는 것이 정확도를 개선하는데 더 효과적임을 알 수 있다.

5. 결 론

일배체형 조합 문제는 생물정보학 분야에서 중요한 문제 중 하나이다. 본 논문에서는 일배체형 조합 문제를 보다 효율적으로 해결하기 위해 WMLF 모델에 유전자형 정보를 도입한 WMLF/GI 모델과 MCIH 모델을 비교, 분석하였다. 현재 유전자형을 도입한 두 모델 사이의 성능에 대한 종합적인 비교가 이루어져 있지 않다. 따라서 본 논문에서는 두 모델의 성능 비교를 위해 유전자형 내의 동형의 분포율이라는 매개변수와 신뢰도의 차이로 대변되는 유전자 서열판독기계의 성능이라는 가정을 이용하였다. 이러한 두 가지 조건하에 새로 설계한 신경망과 기존에 제시된 유전자 알고리즘을 이용하여 일배체형 조합문제의 정확도를 비교, 분석하였다. MCIH 모델 보다 신뢰도를 사용한 WMLF/GI 모델이 일배체형 조합문제에서 전체적으로 높은 정확도를 보였다. 유전자형 정보가 일배체형 조합의 정확도에 미치는 영향은 동형의 비율에 크게 좌우되며, 동형의 비율이 낮으면 상대적으로 신뢰도의 영향력이 커지므로 서열판독기계의 성능이 일배체형 결정에 큰 영향을 미쳤다.

일배체형 조합 문제는 복잡도면에서 어려운 문제로서 앞으로도 이를 해결할 여러 접근 방법에 대한 연구가 요구된다. 또한 실제 현장에서 제기되는 문제들을 반영한 새로운 모델과 문제의 개발도 필요하다.

참 고 문 헌

[1] 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.

[2] 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.

[3] 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.

[4] D. E. Goldberg, *Genetic Algorithms in serarch, Optimization and Machine Learning*, Addison-Wesley, 1989.

[5] 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.

[6] 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.

[7] J. C. Stephens, et al, " Haplotype variation and linkage disequilibrium in 313 human genes," *Science*, Vol.293, pp. 489-493, 2001.

[8] 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.

[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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 강승호, 정인선, 최문호, 임형석, "신뢰도를 가진 SNP 단편들과 유전자형으로부터 일배체형 조합", 정보과학회논문지, 제35권 제11호, pp.509-516, 2008.

정 인 선

e-mail : isjung0@hotmail.com
2001년 여수대학교 전산학과(학사)
2006년 전남대학교 전산학과(석사)
2006년~현 재 전남대학교 전산학과 박사과정
관심분야 : 생물정보학, 알고리즘, 인공지능 등

강 승 호

e-mail : kinston@gmail.com
1994년 전남대학교 전산학과(학사)
2003년 전남대학교 전산학과(석사)
2003년~현 재 전남대학교 전산학과 박사과정
관심분야 : 생물정보학, 알고리즘, 인공지능 등

임 형 석

e-mail : hslim@chonnam.ac.kr
1983년 서울대학교 컴퓨터공학과(학사)
1985년 한국과학기술원 전산학과(석사)
1993년 한국과학기술원 전산학과(박사)
1996년~1997년 미국 Purdue대학 방문교수
1987년~현 재 전남대학교 전자컴퓨터공학부 교수
관심분야 : 알고리즘, 그래프이론, 생물정보학 등