한국지능정보시스템학회논문지 제13권 제2호 2007년 6월(pp.55~68)

품질 정보와 퍼지 추론 기법을 이용한 DNA 염기 서열 배치 알고리즘

김광백
신라대학교 컴퓨터정보공학부
(gbkim@silla.ac.kr)

분자 생물학(computational molecular biology) 분야에서 DNA 염기 서열 배치 알고리즘은 다양한 방법으로 개선되어 왔다. 본 논문에서는 기존의 DNA 염기의 품질 정보(quality information)를 이용한 DNA 염기 서열 배치 방법을 개선하기 위하여 퍼지 논리 시스템(fuzzy logic system)과 DNA 염기 서열 단편의 특징을 적용한 품질 정보와 퍼지 추론 기법을 이용한 DNA 염기 서열 배치 알고리즘을 제안한다. 기존의 알고리즘은 Needleman-Wunsch가 제안한 전역 배치 알고리즘에 각 DNA 염기의 품질 정보를 적용하여 DNA 염기 서열 배치 점수를 계산하였다. 그러나 전체 DNA 염기의 품질 정보를 이용하여 계산하기 때문에 DNA 염기 말단 부분의 품질이 낮은 경우에는 DNA 염기 서열 배치 점수를 계산하는 과정에서 오차가 발생한다. 본 논문에서는 기존의 품질 정보를 이용한 알고리즘을 개선하여 DNA 염기 서열의 말단 부위의 품질이 낮은 경우에도 정확히 서열을 배치할 수 있도록 한다. 또한 DNA 염기 서열 단편의 길이와 낮은 품질의 DNA 염기 빈도를 퍼지 논리 시스템에 적용하여 DNA 염기 서열 배치 점수를 계산하는데 적용되는 매핑 점수 인자(parameter)를 동적으로 조정한다. 제안된 알고리즘의 성능 평가를 위해 NCBI(National Center for Biotechnology Information)의 실체 유전체 데이터를 받아 성능을 분석한 결과, 제안된 알고리즘이 기존의 품질 정보만을 이용한 알고리즘 보다 DNA 염기 서열 배치에 있어서 효율적임을 확인하였다.

논문접수일 : 2006년 12월 게재확정일 : 2007년 05월 교신저자 : 김광백

1. 서론

유전자의 전체 DNA 염기 서열을 얻기 위한 contig 구성 작업에서 DNA 염기 서열 배치는 분자 생물학 분야에서 매우 중요하다[1-5]. 최근에는 자동화된 DNA 염기 서열 분석 장치(automatic sequence analyzer)를 이용하여 DNA 염기 서열 분석 작업에서 요구되는 인적 노력과 소요 시간을 줄일 수 있게 되었지만, DNA 염기 서열의 길이가 매우 긴 특정 유전자의 전체 DNA 염기 서열은 한 번의 실험만으로 모두 해독 할 수 있는 방법은 없다. 이러한 특정 유전자는 여러 개의 단편(fragment)으로 구분한 후에 각 단편들의 DNA 염기 서열을 분석한다. 그리고 이 단편들의 정보를 이용하여 전체 DNA 염기 서열을 재구성해야 하며, 이러한 작업을 config 구성 작업이라 한다[6]. 독립된 실험에서 한번의 분석 작업으로 해독 할 수 있는 길이는 일반적으로 수백 염기쌍(nucleotide base pair)을 넘지 못한다. 특히 DNA 염기 서열 분석 장치를 이용할 때, 동시에 다수의 시료를 분

석할 수 있어 짧은 시간 내에 많은 결과를 얻을 수 있으나 각 DNA 염기 단편들의 말단에 낮은 품질의 DNA 염기가 발견되는 특징이 있다.

생물들의 DNA 염기 서열을 밝혀내는 시퀀싱(sequencing) 작업에 사용되는 PHRED[7] 같은 DNA 염기 결정 프로그램에서 생성되는 품질 정보를 이용한 기존의 알고리즘[8]은 DNA 염기 서열의 말단 부분에 낮은 품질의 DNA 염기가 존재하는 경우에는 DNA 염기 서열 배치 점수를 계산하는데 있어서 오차가 크게 발생한다.

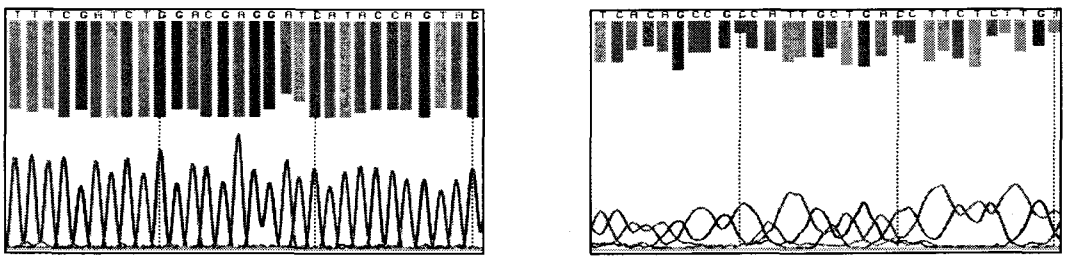
본 논문에서는 기존의 DNA 염기 서열 배치 알고리즘을 개선하기 위하여 DNA 염기 서열 단편의 말단에 나타나는 낮은 품질 정보와 DNA 염기 서열 배치 점수를 계산하는데 사용되는 매핑 점수 인자를 퍼지 논리 시스템에 적용하여 DNA 염기 서열의 말단 부분에 낮은 품질의 DNA 염기가 존재하여도 DNA 염기 서열 배치를 효율적으로 할 수 있는 알고리즘을 제안한다.

2. 품질 정보

DNA 염기 결정 프로그램은 Trace 데이터를 읽어서 DNA 염기들의 서열을 생성하고, 각 DNA 염기에 해당하는 품질 정보를 생성한다. 대부분의 DNA 염기 결정 프로그램이 비슷한 데이터를 생성하므로, 본 논문에서는 잘 알려진 DNA 염기 결정 프로그램 중에서 PHRED를 기준으로 생성된 품질 정보를 이용한다.

PHRED에서는 시퀀싱 머신(sequencing machine)이 크로마토그램(chromatogram)의 정점(peak)을 분석하여 Trace 데이터를 생성한다. 이 데이터는 DNA 염기의 서열인 Fasta와 각 DNA 염기의 품질 정보인 Quality의 두 가지 파일을 생성한다. [그림 1]은 Trace 데이터의 예이다.

이상적인 Trace 데이터는 모든 정점이 동일한 거리를 두고 떨어져 있고, 서로 겹침이 없다. 이러한 데이터에서는 DNA 염기 서열을 정확히 생성할 수 있다. [그림 1]의 (a)는 거의 이상적인 Trace 데이터의 예이다. 정점들 사이의 거리들이 거의 동일하고, 어떤 위치에서 한 곡선의 정점이 다른 세 곡선의 정점보다 훨씬 높다. 그러나 [그림 1]의 (b)는 거리에 따라서 정점이 불분명하게 나타나고 있다. 이러한 이유는 Trace 데이터를 생성할 때 원시 데이터를 생성하는 실험 자체에 오류들이 존재하기 때문이다. 품질이 좋지 않은 Trace 데이터의 특



(a) 품질이 좋은 Trace 데이터 (b) 품질이 좋지 않은 Trace 데이터

A의 신호 G의 신호

C의 신호 T의 신호

[그림 1] Trace 데이터의 예

징은 다음과 같다.

① 두 정점들 사이의 거리가 일정하지 않고 다양하게 분포되어 있다.
② 둘 이상의 곡선이 비슷한 정점을 가지고 있다.
③ 네 곡선의 정점이 모두 매우 낮은 경우가 있다.

이러한 특징 때문에 우리는 그 위치의 DNA 염기가 무엇인지 확신 할 수 없게 되어 낮은 품질의 Trace 데이터를 얻게 된다.

[그림 1]의 윗부분에 있는 문자들은 PHRED가 Trace 데이터를 읽어 들여서 생성한 DNA 염기 서열이고 이를 Fasta라 한다. PHRED는 DNA 염기들의 서열과 함께 품질 점수라 불리는 수들의 서열을 생성하고 이를 Quality라 한다. PHRED의 품질 점수는 0~99 사이의 값을 가지는데, 그 위치의 DNA 염기는 실제와 차이를 나타내는 확률과 관계가 있다. NCBI Handbook에 따르면 PHRED에서 품질 점수는 $10^{-P/10}$의 에러확률에 상응한다. 따라서 품질점수를 Q라 하고 DNA 염기의 에러 확률을 P라 하면 $Q = -10 \times \log_{10} P$인 관계가 성립한다. 예를 들어 해당 위치의 염기가 C이고 품질 점수가 20이라면, 그 위치의 염기 C가 아닐 확률은 0.01이며, 품질 점수가 10이라면 에러 확률은 0.1이 된다. <표 1>은 품질 점수에 따른 에러 확률을 나타낸다.

PHRED는 염기 서열과 품질 점수 생성 후에 서열을 정돈(trimming)하는 과정을 수행한다. 보통 Trace data의 첫 부분과 끝 부분은 실험적 한계로 인해 많은 오류들이 포함되어 있으므로 낮은 품질 점수를 갖는다. 많은 오류를 포함하는 단편의 말단 부분은 실험에 좋지 않은 영향을 끼치기 때문에 제거한다. 따라서 실제 사용되는 서열에서 10보다 작은 품질 점수를 가지는 DNA 염기의 비율은 2~5% 정도이다.

<표 1> 품질 점수에 따른 에러 확률

품질 점수 (Q)	에러 확률 (P)	품질 점수 (Q)	에러 확률 (P)
10	10^{-1}	60	10^{-6}
20	10^{-2}	70	10^{-7}
30	10^{-3}	80	10^{-8}
40	10^{-4}	90	10^{-9}
50	10^{-5}	100	10^{-10}

3. 제안된 DNA 염기 서열 배치 알고리즘

최근에는 DNA 염기 서열을 해독하는 과정에서 shotgun sequencing과 같은 소수의 실험을 제외하고는 동시에 여러 개의 단편을 조립할 필요가 없다. DNA 염기 서열 해독 작업에는 PCR 및 direct sequencing 방법들이 자주 사용되며, 실험자가 적절한 primer를 설계함에 따라 각 단편들의 정보를 순서대로 배치할 수 있다. 서열 배치는 크게 전역 배치(global alignment)와 지역 배치(local alignment)가 있으며, 서열의 최적 배치를 탐색하기 위해, 그 동안 다양한 알고리즘들이 제안되었다.

본 논문에서는 기존의 DNA 염기 서열 배치 알고리즘을 개선하기 위하여 DNA 염기 서열 단편의 말단에 나타나는 낮은 품질 정보와 DNA 염기 서열 배치 점수를 계산하는데 사용되는 매핑 점수 인자를 퍼지 논리 시스템에 적용하여 동적으로 조정하는 알고리즘을 제안한다.

3.1 품질 정보를 적용한 제안된 DNA 염기 서열 배치 알고리즘

3.1.1 DNA 염기 서열

DNA 염기 서열이란 알파벳 Σ에 속한 DNA 염

기들의 나열이다. 본 논문에서는 DNA 염기 서열에 대해 다루고 있으므로 Σ를 {a, g, c, t}로 정의한다. 공백(space)은 $\Delta \notin \Sigma$로 정의한다. DNA 염기 서열 A의 i번째 DNA 염기는 A_i, 부분 DNA 염기 서열 $A_iA_{i+1}\cdots A_j$는 $A[i\cdots j]$로 정의한다.

길이가 각각 m과 n인 두 DNA 염기 서열 $A=A_1A_2\cdots A_m$과 $B=B_1B_2\cdots B_n$이 주어졌을 때, 두 DNA 염기 서열의 배치는 $A^*=A_1^*A_2^*\cdots A_m^*$와 $B^*=B_1^*B_2^*\cdots B_n^*$ $(n, m \le l)$이다. A_i^*와 B_i^*는 그 DNA 염기가 무엇인지에 따라서 다음 세 종류의 매핑(mapping) 중 하나로 분류된다.

- 일치(match) : $A_i^* \neq \Delta$, $B_i^* \neq \Delta$이고, $A_i^* = B_i^*$인 경우.
- 불일치(mismatch) : $A_i^* \neq \Delta$, $B_i^* \neq \Delta$이고, $A_i^* \neq B_i^*$인 경우.
- 갭(gap) : A_i^* 또는 B_i^*가 Δ인 경우.

$A_i^* = B_i^* = \Delta$인 경우는 허용되지 않는다. 각 매핑은 해당되는 점수를 가지는데, 일치는 γ, 불일치는 δ, 갭은 μ를 가진다. 이 γ, δ, μ를 매핑 점수 인자라 부르는데, 응용(application)에 따라 다양한 값을 가진다. 일반적으로 γ는 양수이고, δ와 μ는 음수이다.

3.1.2 품질 정보를 적용한 제안된 DNA 염기 서열 배치 알고리즘

품질 정보를 가지는 DNA 염기 서열 $A=A_1A_2\cdots A_m$은 각 A_i가 Σ의 DNA 염기 중 하나이고 그 문자의 품질 점수가 Q_{A_i}인 DNA 염기 서열이다. 품질 점수 Q_{A_i}는 A_i의 오류 확률이 $10^{-Q_A/10}$이라는 것을 의미한다. 앞으로 품질 정보를 가지는 DNA 염기 서열을 품질 DNA 염기 서열이라 정의하고, 이와 구별하여 품질 정보를 가지지 않는 서열을 일반 DNA 염기 서열로 정의한다.

DNA 염기의 품질 점수의 의미를 좀 더 자세히 살펴보면, DNA 염기 서열에서 어떤 위치의 DNA 염기가 $x \in \Sigma$이고, 그 DNA 염기의 품질 점수가 Q_x라고 하면 그 위치에는 $1-10^{-Q_x/10}$의 확률로 x가 나타난다. 그리고 다른 DNA 염기는 $10^{-Q_x/10}$ 확률이 나타나거나 혹은 공백으로 남는다. 품질 DNA 염기 서열에서 DNA 염기가 없어서 생기는 공백을 '-'로 표시하고, x는 그 위치에서의 대표 DNA 염기이다. 본 논문에서는 다음과 같이 가정한다.

가정 1 : 품질 서열의 위치 i에서 대표 DNA 염기가 나타날 확률은 보통 0.9보다 크다.

가정 2 : 대표 DNA 염기 이외의 DNA 염기들과 공백(-)이 나타날 확률은 모두 같다. 만약 대표 DNA 염기의 품질 점수가 10이면 이 위치의 에러 확률은 0.1이며 대표 DNA 염기가 나타날 확률은 0.9이다. 따라서 대표 DNA 염기 이외의 DNA 염기들과 공백이 나타날 확률은 모두 0.025이다.

전역 배치 길이가 각각 m과 n인 두 DNA 염기 서열 $A=A_1A_2\cdots A_m$과 $B=B_1B_2\cdots B_n$이 주어졌을 때, 두 품질 DNA 염기 서열의 배치는 $A^*=A_1^*A_2^*\cdots A_m^*$와 $B^*=B_1^*B_2^*\cdots B_n^*$ $(n, m \le l)$이다. A_i^*와 B_i^*는 각각 A와 B의 DNA 염기들 사이에 0개 또는 1개 이상의 공백(Δ)을 삽입함으로써 배치된다. 공백(Δ)을 삽입해서 배치되는 것은 일반 DNA 염기 서열과 같다. 배치에 삽입되는 공백(Δ)은 그 부분에 항상 어떤 문자도 존재하지 않

는다는 것을 의미하므로, 그 위치에 DNA 염기 $x \in \Sigma$가 나타날 확률은 0이고 공백(-)이 나타날 확률은 1이다. DNA 염기쌍 A_i^*와 B_i^*는 그 대표 DNA 염기에 따라서 다음과 같이 세 종류의 매핑 중 하나로 분류된다.

- 정규일치(regular-match) : $A_i^* \neq \Delta$, $B_i^* \neq \Delta$ 이고, $A_i^* = B_i^*$인 경우.
- 정규불일치(regular-mismatch) : $A_i^* \neq \Delta$, $B_i^* \neq \Delta$이고, $A_i^* \neq B_i^*$인 경우.
- 정규갭(regular-gap) : A_i^* 또는 B_i^*가 Δ인 경우.

$A_i^* = B_i^* = \Delta$인 경우는 허용하지 않는다. 위 세 매핑은 품질 매핑이고, 일반 DNA 염기 서열의 일치, 불일치, 갭은 일반 매핑이다.

DNA 염기쌍 A_i^*와 B_i^*의 매핑 점수 $S(A_i^*, B_i^*)$는 일반 매핑 점수의 기대 값으로 정의된다. A_i^*와 B_i^*의 실제 DNA 염기 종류에 따라 품질 매핑은 일반 매핑인 일치, 불일치, 갭 중의 하나가 된다. <표 2>는 A_i^*와 B_i^*가 실제 DNA 염기에 따라 일반 매핑으로 분석되는 결과를 나타내었다.

<표 2> 실제 DNA 염기에 따른 일반 매핑

B_i^* \ A_i^*	a	c	t	g	-
a	M	N	N	N	G
c	N	M	N	N	G
t	N	N	M	N	G
g	N	N	N	M	G
-	G	G	G	G	E

- M : 실제 DNA 염기가 같은 경우이다. 따라서 이 경우는 일치 점수 γ를 가진다.
- N : 둘 다 공백이 아니면서 실제 DNA 염기가 서로 다른 경우이다. 따라서 불일치 점수 δ를 가진다.
- G : 한쪽은Σ 이고 다른 한쪽은 공백 (-)인 경우이다. 이 경우는 갭 매핑으로 간주되고 점수 μ를 부여한다.
- E : A_i^*와 B_i^*가 모두 공백인 경우이다. 이 경우엔 이 매핑이 배치 상에서 없는 것으로 간주할 수 있으므로 배치 점수에 영향을 주지 않도록 0의 점수를 부여한다.

따라서 일치 매핑이 될 확률은 $P_m(A_i^*, B_i^*)$, 불일치 매핑이 될 확률을 $P_n(A_i^*, B_i^*)$, 갭 매핑이 될 확률을 $P_g(A_i^*, B_i^*)$라 하면 식 (1)과 같이 나타낼 수 있다.

$$S(A_i^*, B_i^*) = \gamma \times P_m(A_i^*, B_i^*) + \delta \times P_n(A_i^*, B_i^*) + \mu \times P_g(A_i^*, B_i^*) \quad (1)$$

A_i^*의 실제 문자가 $x \in \Sigma$일 확률을 α_x, 공백으로 남을 확률을 α_-, B_i^*의 실제 문자가 $x \in \Sigma$일 확률을 β_x, 공백으로 남을 확률을 β_-로 정의한다.

- 정규일치(regular-match)의 경우

A_i^*와 B_i^*의 대표 DNA 염기를 α라 하면 '가정 2'로 부터 식 (2)를 유도할 수 있다.

$$\frac{1-\alpha_a}{4} = \alpha_c = \alpha_g = \alpha_t = \alpha_-,$$
$$\frac{1-\beta_a}{4} = \beta_c = \beta_g = \beta_t = \beta_- \quad (2)$$

<표 3>은 정규 일치의 확률 표이다.

<표 3> 정규 일치의 확률

B_i^* \ A_i^*	a	c	t	g	-
a	$\alpha_a \beta_a$	X	X	X	X
c	Y	Z	Z	Z	Z
t	Y	Z	Z	Z	Z
g	Y	Z	Z	Z	Z
-	Y	Z	Z	Z	Z

<표 3>에서 각 DNA 염기의 확률은 식 (3)과 같다.

$$a = \alpha$$
$$\{c, g, t\} = \frac{1-\alpha_a}{4} \quad (3)$$

<표 3>의 확률 표에서 나타나는 X, Y, Z는 식 (4)와 같다.

$$X = \frac{(1-\alpha_a)\beta_a}{4}, \quad Y = \frac{\alpha_a(1-\beta_a)}{4},$$
$$Z = \frac{(1-\alpha_a)(1-\beta_a)}{16} \quad (4)$$

<표 3>의 확률 표로부터 식 (5)를 유도할 수 있다. 이 식에서 X와 Y의 값은 퍼지 추론 규칙에 의해 결정된 매핑 점수 인자에 많은 영향을 받는다. 따라서 점화식의 정확성을 향상 시키기 위하여 이 두 부분을 점화식에 추가한다. 그러나 Z는 α_a와 β_a의 품질 점수가 1이라도 0.0056의 매우 작은 값이기 때문에 식 (5)에서는 생략한다.

$$P_m(A_i^*, B_i^*) = \alpha_a \beta_a + 3Z \approx \alpha_a \beta_a$$
$$P_n(A_i^*, B_i^*) = 3X + 3Y + 6Z \approx \frac{\alpha_a + \beta_a - 2\alpha_a \beta_a}{4} \times 3$$
$$P_g(A_i^*, B_i^*) = X + Y + 6Z \approx \frac{\alpha_a + \beta_a - 2\alpha_a \beta_a}{4} \quad (5)$$

그러므로 정규 일치의 매핑 점수는 식 (6)과 같이 나타낼 수 있다.

$$S(A_i^*, B_i^*) = \gamma \times \alpha_a \beta_a + \delta \times \frac{\alpha_a + \beta_a - 2\alpha_a \beta_a}{4} \times 3 + \mu \times \frac{\alpha_a + \beta_a - 2\alpha_a \beta_a}{4} \quad (6)$$

• 정규불일치(regular-mismatch)의 경우

A_i^*의 대표 DNA 염기를 c, B_i^*의 대표 DNA 염기를 a라 하면 '가정 2'로 부터 식 (7)를 유도할 수 있다.

$$\frac{1-\alpha_c}{4} = \alpha_a = \alpha_g = \alpha_t = \alpha_-,$$
$$\frac{1-\beta_a}{4} = \beta_c = \beta_g = \beta_t = \beta_- \quad (7)$$

<표 4>는 정규 불일치의 확률 표이다.

<표 4> 정규 불일치의 확률

B_i^* \ A_i^*	a	c	t	g	-
a	X	$\alpha_c \beta_a$	X	X	X
c	Z	Y	Z	Z	Z
t	Z	Y	Z	Z	Z
g	Z	Y	Z	Z	Z
-	Z	Y	Z	Z	Z

여기서 각 DNA 염기의 확률은 식 (8)과 같다.

$$c = \alpha_c,$$
$$\{a, g, t\} = \frac{1-\alpha_c}{4} \quad (8)$$

<표 4>의 확률 표에서 나타나는 X, Y, Z는 식

(9)와 같다.

$$X=\frac{(1-\alpha_c)\beta_a}{4},\ Y=\frac{\alpha_c(1-\beta_a)}{4},$$
$$Z=\frac{(1-\alpha_c)(1-\beta_a)}{16} \quad (9)$$

<표 4>의 확률 표로부터 식 (10)을 유도할 수 있다. 식 (10)에서 X와 Y의 값은 퍼지 추론 규칙에 의해 결정된 매핑 점수 인자에 많은 영향을 받기 때문에 점화식의 정확성을 향상 시키기 위하여 이 두 부분을 점화식에 추가한다. 그러나 Z는 α_a와 β_a의 품질 점수가 1이라도 0.0056의 매우 작은 값이기 때문에 식 (10)에서 생략한다.

$$P_m(A_i^*, B_i^*)=X+Y+2Z\approx\frac{\alpha_c+\beta_a-2\alpha_c\beta_a}{4}$$
$$P_n(A_i^*, B_i^*)=\alpha_c\beta_a+2X+2Y+7Z\approx\frac{\alpha_c+\beta_a}{2}$$
$$P_g(A_i^*, B_i^*)=X+Y+6Z\approx\frac{\alpha_c+\beta_a-2\alpha_c\beta_a}{4} \quad (10)$$

그러므로 정규 일치의 매핑 점수는 식 (11)과 같이 나타낼 수 있다.

$$S(A_i^*, B_i^*)=\gamma\times\frac{\alpha_c+\beta_a-2\alpha_c\beta_a}{4}+\delta\times\frac{\alpha_c+\beta_a}{2}$$
$$+\mu\times\frac{\alpha_c+\beta_a-2\alpha_c\beta_a}{4} \quad (11)$$

• 정규 갭(regular-gap)의 경우

A_i^*의 대표 DNA 염기를 a, B_i^*의 대표 DNA 염기를 공백(Δ)이라 정의하면 B_i^*가 공백(Δ)이 되므로 $\beta_-=1$이다. 정규 갭의 매핑 점수는 식 (12)와 같다.

$$S(A_i^*, B_i^*)=\mu\times(1-\alpha_-)=\mu\times\frac{3+\alpha_a}{4} \quad (12)$$

일반 DNA 염기 서열과 같이 품질 DNA 염기 서열의 배치 점수 $S(A_i^*, B_i^*)$는 배치를 이루는 모든 DNA 염기 쌍들의 매핑 점수의 합으로 정의된다. 즉, 식 (13)과 같이 정의된다.

$$S(A_i^*, B_i^*)=\sum_{i=1}^{l}S(A_i^*, B_i^*) \quad (13)$$

본 논문에서 제안된 알고리즘은 길이가 각각 m과 n인 두 DNA 염기 서열이 주어 졌을 때, DNA 염기 서열의 배치 점수가 가장 높은 경우에 최적 배치로 간주하고 최적 배치를 탐색한다. $H_{i,j}$를 $A[1\cdots i]$와 $B[1\cdots j]$의 최적 배치 점수라 할 때, $H_{i,j}$는 동적 프로그래밍 기법을 적용하여 계산할 수 있고, 이 기법은 기존의 Needleman-Wunsch 알고리즘과 같은 구조를 가진다. 따라서 $O(mn)$의 메모리와 $O(mn)$의 시간이 소요된다. [그림 2]는 품질 DNA 염기 서열을 배치하기 위한 점화식이다.

$\gamma>0$: 일치 점수
$\delta<0$: 불일치 점수
$\mu<0$: 갭 점수
Q_x : 문자 x의 품질 점검

$$P_x=1-10^{-Q_x/10}$$

$$S(A_i, B_j)=\begin{cases}\gamma\times P_{A_i}P_{B_j}+\delta\times\dfrac{3(P_{A_i}+P_{B_j}-2P_{A_i}P_{B_j})}{4} \\ +\mu\times\dfrac{P_{A_i}+P_{B_j}-2P_{A_i}P_{B_j}}{4} \\ \gamma\times\dfrac{P_{A_i}+P_{B_j}-2P_{A_i}P_{B_j}}{4}+\delta\times\dfrac{P_{A_i}+P_{B_j}}{2} \\ +\mu\times\dfrac{P_{A_i}+P_{B_j}-2P_{A_i}P_{B_j}}{4}\end{cases}$$

$$H_{0,0}=0,\ if\ i=0\ or j=0$$

$$H_{i,j}=max\begin{cases} H_{i-1,j-1}+S(A_i,B_j) \\ H_{i-1,j}+\mu\times\dfrac{3+P_{A_i}}{4} \\ H_{i,j-1}+\mu\times\dfrac{3+P_{B_j}}{4} \end{cases}$$

$$(1\le i\le m, 1\le j\le n)$$

[그림 2] 품질 서열 배치를 위한 점화식

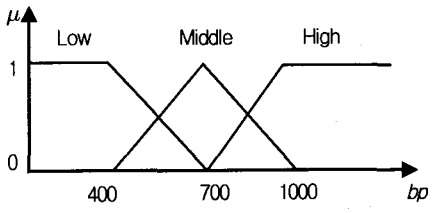
3.2 퍼지 추론 규칙을 이용한 매핑 점수 인자 조정

DNA 염기 서열을 해독하는 과정에서 단편의 말단 부분이 실험적인 한계로 인해 20이하의 품질이 나타나는 경우가 발생한다. 본 논문에서는 이러한 품질을 가지는 DNA 염기를 낮은 품질로 정의한다. 품질 정보를 이용한 기존의 알고리즘[8]은 매핑 점수 인자를 사용자가 입력 받아서 최적 단편을 계산하기 때문에 각 DNA 염기들의 길이의 차이가 크고, 서열 단편의 말단 부분의 DNA 염기 품질이 좋지 않은 경우에는 서열 배치 점수를 계산하는데 있어서 오차가 자주 발생한다. 본 논문에서는 이러한 문제점을 개선하기 위해 퍼지 추론 규칙에 각 DNA 염기 단편의 길이와 품질 정보가 낮은 단편의 정보를 적용하여 개선한다. 즉 매핑 점수 인자를 퍼지 논리 시스템에 적용하여 동적으로 조정한다. 퍼지 논리 시스템의 입력은 각 DNA 염기 단편의 길이와 단편의 품질이 좋지 않은 단편의 빈도 수이고, 출력은 불일치 매핑 점수 인자이다. 그리고 최종 매핑 점수는 불일치 점수를 기준으로 일치 점수와 갭 점수를 임계 범위로 설정한다. 퍼지 논리 시스템은 입력 신호의 퍼지화, 전문가의 지식에 기반을 둔 퍼지 규칙에 의한 퍼지 추론, 비 퍼지화로 구성된다. 퍼지 규칙을 추론하기 위해 Max-Min 추론을 적용한다. 비 퍼지화기는 퍼지 추론의 결과인 퍼지 값을 단일 실수 값으로 변화시키는 부분으로 본 논문에서는 식 (14)와 같은 무게 중심법[9]을 적용한다.

$$y^*=\frac{\sum\mu(y_i)x_i}{\sum\mu(y_i)} \qquad (14)$$

3.2.1 각 단편의 길이에 대한 소속 함수

각 단편의 길이에 대한 소속 함수를 [그림 3]과 같이 설계하고 소속도를 계산한다. 이때, Low 구간은 단편의 길이가 짧은 구간이고, Middle 구간은 단편의 길이가 중간 정도인 구간이고, High 구간은 단편의 길이가 긴 구간이다. PHRED에서 단편의 길이를 추출할 때에 실험적 한계로 인해 DNA 염기 단편의 길이가 1000개 이상을 가지는 경우는 매우 드물다. 따라서 본 논문에서 단편의 길이에 대한 퍼지 소속 구간은 다양한 실험을 기반으로 [그림 3]과 같이 Low 구간은 [0, 700], Middle 구간은 [400, 1000], High 구간은 [700, ∞]로 설정하였다.



[그림 3] 단편의 길이 소속 함수

<표 5> 단편 길이의 퍼지 값

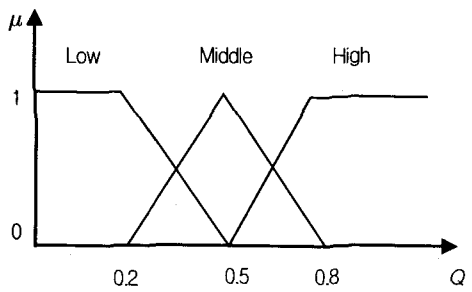
퍼지 값 (소속 정도)	소속구간 (각 단편의 길이)
Low	[0, 700]
Middle	[400, 1000]
High	[700, ∞]

3.2.2 낮은 품질 DNA 염기의 빈도 수에 대한 소속 함수

각 단편에서 20이하의 낮은 품질의 빈도수 Q를 식 (15)와 같이 계산한다.

$$Q = \frac{\text{낮은 품질의 단편 개수}}{\text{단편의 전체 길이}} \quad (15)$$

이때, Low 구간은 낮은 품질 DNA 염기의 빈도수가 적은 구간이고, Middle 구간은 낮은 품질 DNA 염기의 빈도수가 중간 정도인 구간이고, High 구간은 낮은 품질 DNA 염기의 빈도 수가 높은 구간이다. [그림 4]는 낮은 품질 DNA 염기의 빈도 수에 대한 소속 함수이다. [그림 4]에서 낮은 품질 DNA 염기의 빈도 수에 대한 소속 함수는 다양한 실험을 통하여 Low 구간은 [0, 0.5], Middle 구간은 [0.2, 0.8], High 구간은 [0.5, 1] 으로 설정하였다.



[그림 4] 낮은 품질 DNA 염기의 빈도 수 소속 함수

<표 6> 낮은 품질 DNA 염기 빈도의 퍼지 값

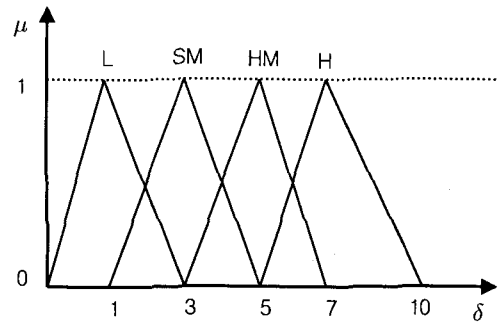
퍼지 값 (소속 정도)	소속구간 (낮은 품질의 빈도수)
Low	[0, 0.2]
Middle	[0.2, 0.8]
High	[0.5, 1]

3.2.3 불일치 매핑 점수 인자에 대한 출력 소속 함수

단편의 길이의 소속도와 낮은 품질 DNA 염기 빈도 수의 소속도를 <표 7>과 같은 퍼지 추론 규칙을 적용하여 추론한 후에, 무게 중심 법을 적용하여 비퍼지화 하고 최종 불일치 매핑 점수를 구한다. [그림 5]는 불일치 매핑 점수에 대한 출력 소속 함수이다. [그림 5]에서 불일치 매핑 점수 인자에 대한 출력 소속 함수는 다양한 실험을 기반으로 L 구간은 [0, 3], SM 구간은 [1, 5], HM 구간은 [3, 7], H 구간은 [5, 10]으로 설정하였다.

<표 7> 불일치 매핑 점수에 대한 퍼지 추론 규칙

bp \ Q	Low	Middle	High
Low	L	SM	SM
Middle	SM	SM	HM
High	SM	HM	H

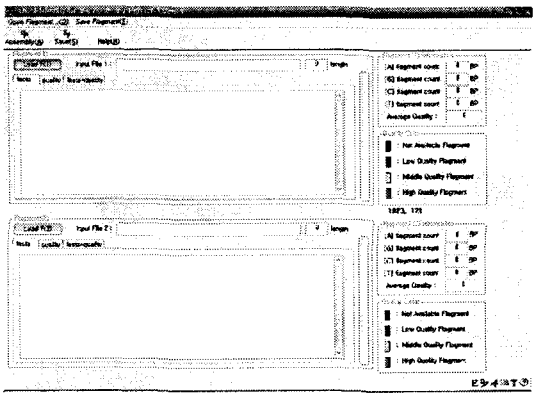


[그림 5] 불일치 매핑 점수 출력 소속 함수

퍼지 논리 시스템에서 출력된 불일치 매핑 점수에 ±1차이로 일치 매핑 점수와 갭 매핑 점수를 각 단편이 바뀔 때마다 동적으로 조정한다.

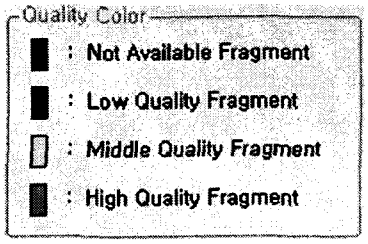
4. 실험 및 결과 분석

본 연구에서의 실험 환경은 삼성 Sens 노트북 x10 (M) 1.3GHz CPU와 512M RAM이 장착된 PC 상에서 VC++ 6.0으로 구현하였다. 테스트 데이터는 NCBI(http://www.ncbi.nlm.nih.gov/Traces/trace.fcgi)에서 실제 유전체(gnome) 데이터를 받아 실험하였다. 실험에 사용된 유전체의 이름은 "gnl|ti|1147316796"이고, "Influenza A Virus"이다. 각 유전체는 PHRED로 생성된 Fasta 파일과 Quality 파일을 가진 서열 쌍 166개이다. 길이는 최소 311bp에서 최대 872bp이다. 제안된 최적 서열 배치를 탐색하는 알고리즘을 구현한 초기 화면은 [그림 6]과 같다.



[그림 6] 제안된 최적 서열 배치 탐색 초기 화면

[그림 6]은 제안된 알고리즘을 이용하여 2개의 DNA 염기 서열 단편을 불러와서 서열 배치를 할 수 있도록 하였고 각 DNA 염기 서열 단편의 Fasta와 Quality를 동시에 불러와서 각 DNA 염기의 품질을 한눈에 볼 수 있도록 나타내었다. 그리고 DNA 염기와 단편을 단편의 품질에 따라 색상을 설정하여 각 DNA 염기들의 품질 정도를 색상으로 확인할 수 있도록 하였다. [그림 7]과 <표 8>은 각 DNA 염기의 품질에 따른 색상과 이 색상이 나타내는 품질의 컬러 범위이다.

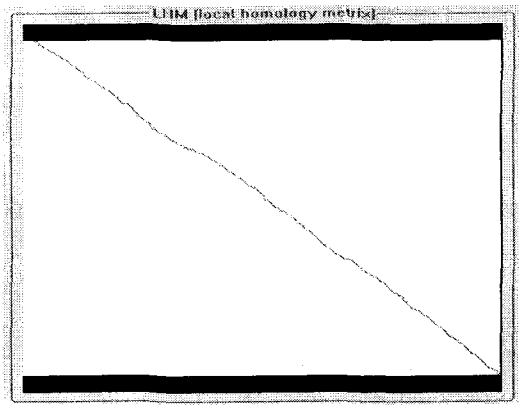


[그림 7] 품질에 따른 색상

<표 8> 품질의 컬러 범위

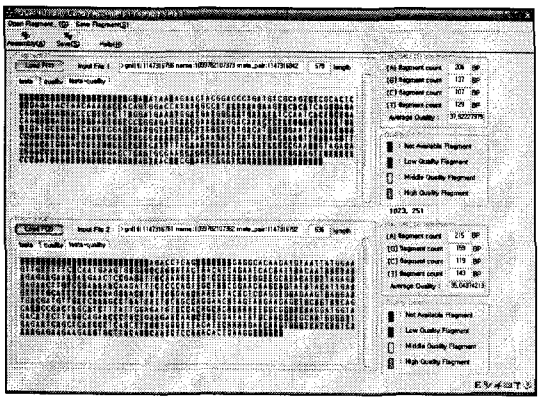
	품질 점수 범위
사용되지 않는 단편	0
낮은 품질의 단편	1 ~ 20
중간 품질의 단편	21 ~ 40
높은 품질의 단편	41 ~ 99

[그림 8]은 제안된 알고리즘에서 DNA 염기 서열의 최적 배치 결과를 표시하기 위해서 LHM(local homology mertrix)에서 추적 경로를 나타내었다. [그림 8]에서 가로와 세로 축은 두 개의 DNA 염기 단편을 각각 배치하여 최적 경로를 그래프로 표시한 것이다.



[그림 8] 두 DNA 염기 서열의 최적 경로 그래프

기존의 품질 정보를 적용한 DNA 염기 서열 배치 알고리즘[7]과 제안된 알고리즘 간의 성능을 비교하기 위해 NCBI에서 받은 실제 DNA 염기 서열 쌍 166개를 실험에 적용하였다. 실험에 적용된 "gnl|ti|1147316796"의 166개 DNA 염기 서열 쌍 중에서 166개 모두가 말단 부분의 DNA 염기의 품질이 20이하인 낮은 품질 단편이 나타났다. [그림 9]는 두 DNA 염기 서열의 말단 부분에서 낮은 품질 정보가 나타난 화면이다.



[그림 9] 두 DNA 염기 서열 단편 화면

기존의 알고리즘[7]과 제안된 알고리즘 간의 DNA 염기 서열의 최적 배치 결과를 비교하기 위해 실험을 통해서 계산된 두 DNA 염기 서열 단편 배치에 대한 일부분을 각각 [그림 10]과 [그림 11]로 나타내었다.

--GTA--GACATA···CGGGTCATAT---TGGTTTTCCTAA
GGGAGGGG-CTA-···---CAC-AATTAACAA-TAA-CTAA

[그림 10] 기존 알고리즘[7]의 최적 배치

--GTA--GACATAAA···TCGGGTCATATTGGTTTT-CCTAA
GGGGAGGGG-CTA-A···T---CAC-AAT----TAACAATAA

[그림 11] 제안된 알고리즘의 최적 배치

기존의 알고리즘[7]과 제안된 알고리즘 간의 최적 DNA 염기 서열 배치 결과는 <표 9>와 같다.

<표 9> 최적 서열 배치 결과

	기존 알고리즘[7]	제안된 알고리즘
일치 개수	12/29	14/29
불일치 개수	12/29	10/29
갭 개수	15/29	15/29

<표 9>에서 알 수 있듯이 제안된 알고리즘이 기존의 알고리즘[7]보다 DNA 염기의 일치 개수가 높게 나타났고 불일치 개수는 적게 나타났다. 따라서 제안된 알고리즘이 기존의 알고리즘으로 DNA 염기 서열 배치 점수를 계산하는 것보다 오류가 적게 발생하여 DNA 염기 서열 배치율이 개선됨을 확인할 수 있다.

<표 10>은 DNA 염기 서열 배치에서 기존의 알고리즘과[7]과 제안된 알고리즘 간의 최적 배치에 대한 일치 및 불일치 개수를 비교한 것이다.

<표 10> 두 알고리즘간의 DNA 염기 서열 배치 일치 및 불일치 결과 비교

	일치	불일치
두 알고리즘 간의 DNA 염기 서열 배치 일치성	12/166	144/166

<표 10>에서 제안된 알고리즘과 기존 알고리즘[7] 간의 염기 서열 배치 결과가 불일치 하는 경우를 분석한 결과, 제안된 알고리즘이 기존의 방법보다 DNA 염기 서열 배치에 대한 정확성이 높게 나타났다. 그 이유는 본 논문에서는 DNA 염기 서열 단편의 길이와 낮은 품질의 DNA 염기 빈도 수를 퍼지 논리 시스템에 적용하여 DNA 염기 서열 배

치 점수를 계산하는데 적용되는 매핑 점수 인자(parameter)를 동적으로 조정하여 말단부분의 품질이 낮은 DNA 염기의 배치 성능을 개선하였기 때문이다. 두 알고리즘에서 DNA 염기 서열 배치가 일치 하는 경우는 단편의 평균 품질이 40이상으로 높게 나타나서 DNA 염기 서열의 배치 점수를 계산하는 과정에서 오류의 차이가 적은 경우이다. 두 알고리즘에서 DNA 염기 서열의 배치가 불일치한 경우는 두 단편의 말단 부위의 품질이 20 이하인 낮은 경우이다.

5. 결론

본 논문에서는 기존의 품질 정보(quality information)를 이용한 DNA 염기 서열 배치 방법을 개선하기 위하여 기존의 알고리즘에서 매핑 점수를 계산하는 점화식에 이용되는 매핑 점수 인자(parameter)를 퍼지 논리 시스템(fuzzy logic system)을 적용하여 동적으로 조정하는 방법을 제안하였다. 기존의 알고리즘은 Needleman-Wunsch가 제안한 전역 배치 알고리즘에 각 DNA 염기의 품질 정보를 적용하여 DNA 염기 서열 배치 점수를 계산하였다. 그러나 전체 DNA 염기의 품질 정보를 이용하여 계산하기 때문에 DNA 염기 말단 부분의 품질이 낮은 경우에는 DNA 염기 서열 배치 점수를 계산하는 과정에서 오차가 발생하였다. 따라서 본 논문에서는 DNA 염기 서열 단편의 길이와 낮은 품질의 DNA 염기 빈도 수를 퍼지 논리 시스템에 적용하여 기존의 전역 배치 알고리즘과 품질 정보만을 이용한 알고리즘 보다 DNA 염기 서열 배치율을 개선하였다.

제안된 방법을 적용하여 최적의 서열 배치를 탐색한 결과, 말단 부분에 품질이 낮은 단편이 있는 경우에도 제안된 방법이 기존의 전역 배치 알고리즘과 품질 정보만을 이용한 알고리즘 보다 DNA 염기 서열 배치율이 개선되었다. 본 논문에서는 전역 배치 알고리즘에 품질 정보와 단편의 매핑 점수 인자를 퍼지 논리 시스템에 적용하여 서열 배치를 탐색하였다.

향후 연구 방향은 지역 배치 알고리즘을 적용하여 서열 배치의 정확성을 높일 수 있도록 할 것이다.

참고문헌

[1] Waterman, M. S., *Introduction to Computational Biology*, Champman and Hall, 1995.

[2] Gusfield, D., *Algorithms on Strings, Trees and Sequences : Computer Science and Computational Biology*, Cambridge University Press, 1997.

[3] Apostolico, A. and R. Giancarlo, "Sequence Alignment in Molecular Biology", *Journal of Computational Biology*, Vol.5 No.2(1998), 173~196.

[4] Pevzner, P., *Computational Molecular Biology : An Algorithmic Approach*, MIT Press, 2000.

[5] Needleman, S. B. and C. D. Wunsch, "A general method applicable to the search for similarities in the aminoacid sequences of two proteins", *Journal of Molecular Biology*, Vol.48(1970), 443~453.

[6] Staden, R. "A new computer method for the storage and manipulation of DNA gel reading data", *Nucleic Acids Res.*, Vol.8(1980), 3673~3694.

[7] Ewing, B., L. Hillier, M. C. Wend, and P. Green, "Base-calling of automated sequen-

cer traces using phred. I. Accuracy assessment", *Genome Research*, Vol.8, No.3(1998), 175~185.

[8] Na, J. C. K. H. Noh, and K. S. Park, "DNA Sequencing Algorithm Using Quality Information", *The Korea Information Science Society*, Vol.32, No.11(2005), 578-586.

[9] George, J. K. and Y. Bo, *Fuzzy Sets and Fuzzy Logic Theory and Applications*, Prentice Hall PTR, 1995.

Abstract

A DNA Sequence Alignment Algorithm Using Quality Information and a Fuzzy Inference Method

Kwang-Baek Kim*

DNA sequence alignment algorithms in computational molecular biology have been improved by diverse methods. In this paper, we proposed a DNA sequence alignment algorithm utilizing quality information and a fuzzy inference method utilizing characteristics of DNA sequence fragments and a fuzzy logic system in order to improve conventional DNA sequence alignment methods using DNA sequence quality information. In conventional algorithms, DNA sequence alignment scores were calculated by the global sequence alignment algorithm proposed by Needleman-Wunsch applying quality information of each DNA fragment. However, there may be errors in the process for calculating DNA sequence alignment scores in case of low quality of DNA fragment tips, because overall DNA sequence quality information are used. In the proposed method, exact DNA sequence alignment can be achieved in spite of low quality of DNA fragment tips by improvement of conventional algorithms using quality information. And also, mapping score parameters used to calculate DNA sequence alignment scores, are dynamically adjusted by the fuzzy logic system utilizing lengths of DNA fragments and frequencies of low quality DNA bases in the fragments. From the experiments by applying real genome data of NCBI (National Center for Biotechnology Information), we could see that the proposed method was more efficient than conventional algorithms using quality information in DNA sequence alignment.

Key words : DNA sequence alignment algorithms, Quality information, Fuzzy inference method, Mapping score parameters, Needleman-Wunsch, NCBI (National Center for Biotechnology Information)

* Division of Computer and Information Engineering, Silla University