

퍼지 논리를 이용한 DNA 염기 서열의 단편 결합

김광백* · 우영운**

1. 서 론

DNA 염기 서열 배치 알고리즘은 분자 생물학 분야에서 단백질과 핵산 서열들의 분석에서 중요한 방법이다. 생물학적인 염기 서열들은 그들 사이의 유사성과 차이점을 나타내기 위해 정렬된다. DNA 염기 서열 배치는 두 개 혹은 그 이상의 DNA 서열을 비교 및 정렬하여 상동성(homology)이 높은 서열들을 알아내 서열의 기능을 유추하거나, 각 서열간의 진화적 연관성이나 관련 기능 등을 예측하기 위한 것이 주목적이다. 다양한 방법으로 많은 알고리즘이 개선되어 왔으며 이는 분자 생물학 분야에서 매우 중요한 작업이며 생체 정보 처리의 기본이라고 할 수 있다.

DNA sequencing 프로젝트에서 매우 긴 DNA 염기 서열을 밝혀내려는 경우, 우선 그 DNA 가닥을 여러 개의 DNA 단편(fragment)들로 만든 후, 각 단편들의 염기 서열을 알아낸다. 그리고 염기 서열이 밝혀진 단편들로부터 본래의 긴 DNA 염기 서열을 재구성한다. 이 경우 발생하는 재구성 문제를 'contig 구성 문제'[1]라 하며, 궁극적으로 이 문

제는 그 자체의 복잡성과 그로 인한 많은 계산량 때문에 컴퓨터의 빠른 계산능력을 필요로 한다. 현재까지 단편 염기 서열로부터 contig를 구성하는 프로그램으로는 SEQAID[2], CAP[3], FAP[4] 등이 알려져 있다. 이들 대부분 프로그램의 입력 단편에 사용할 수 있는 염기는 A,C,G,T이며, 의미가 모호한 염기에 대해서는 N으로 나타낸다.

본 연구에서는 기존의 contig 구성 프로그램의 단점인 결합 실패를 보완하는 알고리즘을 적용한다. 본 연구의 알고리즘은 기존의 일치율만 가지고 sequencing하는 방법에 퍼지 추론 기법[5,6]을 추가하여 결합 실패가 발생하지 않고 모호한 염기에서도 결합이 가능하도록 한다.

본 연구에서는 테스트 데이터를 얻기 위해서 완성된 단백질 지놈인 'Synechocystis PCC6803'을 임의의 mutation을 유발하였다. 이들에 대해 제안된 방법으로 실험한 결과, 모두 original sequence를 구성하였으며, 실행 시간은 단편의 수에 비례하는 것을 확인하였다. 그리고 contig 구성 프로그램의 단점인 결합 실패가 발생하지 않았다.

2. DNA 염기 서열 분석 알고리즘

2.1 DNA 염기 서열 분석 과정

본 연구의 DNA 염기 서열 분석 과정은 기존에

* 교신저자(Corresponding Author) : 김광백, 주소 : 부산광역시 사상구 신라대학길 100번(617-736), 전화 : 051)999-5052, FAX : 051)999-5657, E-mail : gbkim@silla.ac.kr

* 신라대학교 컴퓨터정보공학부 교수

** 동의대학교 멀티미디어공학과 교수
(E-mail : ywwoo@deu.ac.kr)

알려진 자동 염기 서열 분석기를 이용하여 한번에 분석된 약 700개의 단편들을 한 주형으로 만들어 PCR(Polymerase Chain Reaction) 방법[7]을 이용하여 클론을 3개 생성한다. 생성된 클론들을 600~700개의 임계치로 단편화하여 기준 주형과 비교하여 일치율을 측정할 수 있도록 한다. 이 단편 쌍들의 중첩정도를 기준으로 주형마다 2개의 결합 후보 단편을 추출하여 추출된 각 단편들의 일치율과 A,G,C,T 소속도 및 각 A,G,C,T 이전 빈도수를 퍼지 추론 규칙을 이용하여 결합 여부를 판단한다. 결정된 최적의 비교 단편을 기준 단편과 결합하는 과정으로 수행하여 단편이 없을 때까지 반복하여 서열 결합을 완성한다. 그림 1은 본 연구의 DNA 염기 서열 결합 과정이다.

2.2 품질 정보

DNA 염기 결정 프로그램은 Trace 데이터를

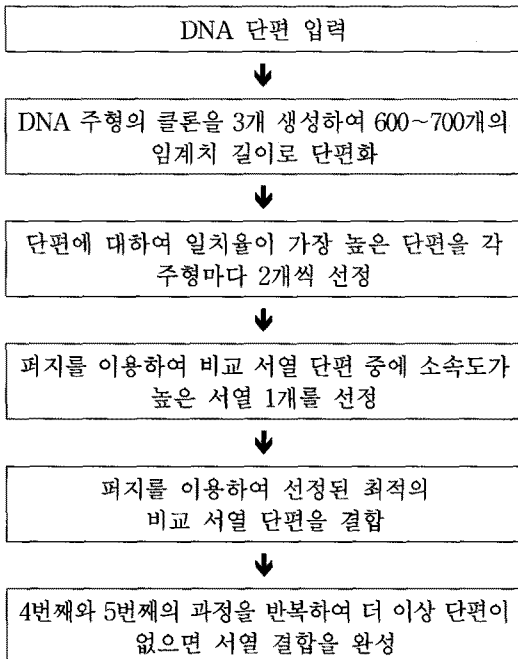


그림 1. 본 연구의 DNA 염기 서열 결합 과정

읽어서 DNA 염기들의 서열과 각 DNA 염기에 해당하는 품질 정보(quality score)를 생성한다. 본 연구에서는 염기 결정 프로그램 중 대표적인 PHRED에서 생성된 품질 정보를 다룬다. PHRED에서는 Sequencing machine이 chromatogram의 정점(peak)을 분석하여 trace data를 생성한다. 이 data는 DNA 염기의 서열인 FASTA와 각 염기의 품질 정보인 Quality의 파일을 생성한다. 이상적인 경우에는 trace data에서 모든 정점이 거의 동일한 거리를 두고, 서로 겹치지 않는다. 그러나 실제 trace data는 실험적인 한계로 인하여 오류들이 존재하기 때문에 이상적인 것과는 상이하다. 좋지 않은 품질의 trace data의 특징은 다음과 같다.

- (1) 두 정점들 사이의 거리가 일정하지 않고 다양하게 분포 되어 있다.
- (2) 둘 이상의 곡선이 비슷한 정점을 가지고 있다.
- (3) 네 곡선의 정점이 모두 매우 낮은 경우가 존재한다.

이러한 특징 때문에 해당 위치의 DNA 염기가 무엇인지 확신 할 수 없게 되어 낮은 품질의 DNA FASTA를 얻게 된다.

2.3 Compute Agreement 알고리즘

본 연구에서는 선택된 기준 단편과 비교 될 단편의 일치율을 계산하기 위하여 Compute Agreement 알고리즘을 제안하여 적용한다. 이 알고리즘은 기준 단편과 비교 단편의 길이만큼 검색하면 일치율과 단편 결합 위치를 측정할 수 있기 때문에 일치율 계산 시간이 감소된다. Compute Agreement 알고리즘은 다음과 같다.

단계 1. 기준 단편을 행으로 설정하고, 비교 단편은 열로 설정 한다.

단계 2. 그림 2와 같이 열의 첫 번째 값과 기준

	C	G	T	C	A	G	A	T	A
C									
A									
G									
A									
T									
A									
G									
T									
C									

그림 2. 제안된 Compute Agreement 알고리즘

단편의 각 염기의 일치 여부를 비교한다. 이 때, 일치하면 대각선으로 탐색하면서 일치 여부를 비교하여 행 또는 열의 마지막까지 모두 일치하면 이 길이를 후보 일치율의 크기로 설정한다.

단계 3. 그림 2와 같이 행의 첫 번째 값과 비교 단편의 각 염기의 일치 여부를 비교한다. 이 때, 일치하면 대각선으로 탐색하면서 일치 여부를 비교하여 행 또는 열의 마지막까지 모두 일치하면 이 길이를 후보 일치율의 크기로 설정한다.

단계 4. 단계 2과 단계 3의 후보 일치율 중에 높은 일치율이 최종 일치율이 되고, 행의 마지막까지 일치하면 왼쪽으로 결합하고, 열의 마지막까지 일치하면 오른쪽으로 결합한다.

본 알고리즘을 이용하여 측정된 각 주형의 비교 단편 중에서 가장 일치율이 높은 2개의 승자 단편을 추출하여 퍼지 추론 기법을 이용하여 결합할 단편을 설정한다.

2.4 퍼지를 이용한 결합할 단편 선정

기존의 SEQAID, CAP, FAP 프로그램에서는 일치율만 결합 여부 판단에 사용하여서 최소한의 오류율을 가지는 단편을 결합하였다. 본 연구에서

제안된 퍼지 추론 기법을 이용하여 일치율을 측정하여 승자 단편 6개의 각 염기의 소속도와 이전 빈도수의 소속도를 퍼지 추론 규칙에 적용하여 결합 여부를 판단한다.

퍼지를 이용하여 결합할 단편을 선정하는 알고리즘은 다음과 같다.

단계 1. 승자 단편의 연장 염기를 각 A,G,C,T별로 소속도를 계산한다. 여기서 퍼지 입력 값은 식 (1)과 같다.

$$\text{각 } A, G, C, T \text{의 퍼지 값} = \frac{\text{각 } A, G, C, T \text{의 수}}{\text{전체 염기수}} \quad (1)$$

단계 2. 이전 빈도수의 소속도를 계산한다. 이 때, 처음 염기의 이전 빈도수는 없으므로 식 (2)와 같이 계산하고, 두 번째 염기부터는 식 (1)과 같이 소속도를 계산한다.

$$\text{각 } A, G, C, T \text{의 처음 염기의 퍼지 값} = \frac{\text{승자 염기수 일치율 합}}{\text{전체 일치율}} \quad (2)$$

단계 3. 각 DNA 염기 A,G,C,T의 소속도와 이전의 빈도수를 제안한 퍼지 추론 규칙을 적용하여 결합 여부를 결정 한다.

위의 단계에서 결정된 최적의 비교 단편을 결합하고, 더 이상 단편이 없을 때까지 반복하여 서열 결합을 완성한다.

2.4.1 각 DNA 염기 A,G,C,T에 대한 소속함수
승자 단편의 연장 염기 A,G,C,T에 대해 그림 3과 같은 소속 함수에 적용하여 소속도를 계산한다.

이때, Low 구간은 기준 단편에 대한 소속 정도가 낮은 구간이고, High 구간은 기준 단편에 대한 소속 정도가 높은 구간이다.

(1) Low 구간

$\mu(L)$ 은 Low 구간의 각 승자 단편의 연장 염기 A,G,C,T의 소속도이다.

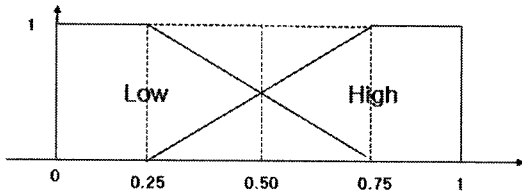


그림 3. 각 A,G,C,T 염기에 대한 소속 함수

$$\begin{aligned} & \text{If}(L \leq 0.25) \text{ then } \mu(L) = 1 \\ & \text{Else If}(L \geq 0.75) \text{ then } \mu(L) = 0 \\ & \text{Else } \mu(L) = \frac{(0.75 - L)}{(0.75 - 0.25)} \end{aligned}$$

(2) High 구간

$\mu(H)$ 는 High 구간의 각 승자 단편의 연장 염기 A,G,C,T의 소속도 이다.

$$\begin{aligned} & \text{If}(H \leq 0.25) \text{ then } \mu(H) = 0 \\ & \text{Else If}(H \geq 0.75) \text{ then } \mu(H) = 1 \\ & \text{Else } \mu(H) = \frac{(H - 0.25)}{(0.75 - 0.25)} \end{aligned}$$

2.4.2 각 DNA 염기 A,G,C,T의 이전 빈도수에 대한 소속 함수

승자 단편의 연장 염기의 이전 빈도수를 그림 4와 같은 소속 함수에 적용하여 각 DNA 염기 A,G,C,T의 이전 빈도수에 대한 소속도를 계산한다. 이때, Low 구간은 기준 단편에 대한 이전 빈도수의 소속 정도가 낮은 구간이고, High 구간은 기준 단편에 대한 이전 빈도수의 소속 정도가 높은 구간이다.

(1) Low 구간

$\mu(L)$ 은 Low구간의 승자 각 단편의 A,G,C,T에 대한 연장 염기의 이전 빈도수의 소속도 이다.

$$\begin{aligned} & \text{If}(L \leq 0.25) \text{ then } \mu(L) = 1 \\ & \text{Else If}(L \geq 0.75) \text{ then } \mu(L) = 0 \\ & \text{Else } \mu(L) = \frac{(0.75 - L)}{(0.75 - 0.25)} \end{aligned}$$

(2) High 구간

$\mu(H)$ 는 High구간의 승자 각 단편의 A,G,C,T

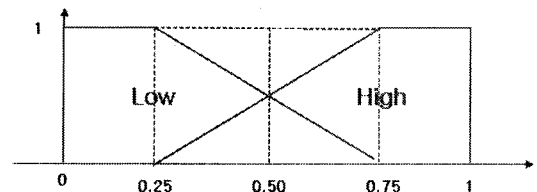


그림 4. 각 A,G,C,T의 이전 빈도수에 대한 소속 함수

에 대한 연장 염기의 이전 빈도수 소속도이다.

$$\begin{aligned} & \text{If}(H \leq 0.25) \text{ then } \mu(H) = 0 \\ & \text{Else If}(H \geq 0.75) \text{ then } \mu(H) = 1 \\ & \text{Else } \mu(H) = \frac{(H - 0.25)}{(0.75 - 0.25)} \end{aligned}$$

2.4.3 단편 결합에 대한 추론 규칙

여기서 $\mu(A), \mu(G), \mu(C), \mu(T)$ 는 각 염기의 소속도이고, $\mu(Y_a), \mu(Y_g), \mu(Y_c), \mu(Y_t)$ 는 각 염기의 이전 빈도수의 소속도 이다. $\mu(W)$ 는 최종적인 각 염기에 대한 결합 여부의 소속도 이다. 4가지 염기에 대해서 $\mu(W)$ 값을 추론하는 규칙은 다음과 같다.

- If $\mu(A)$ is L, $\mu(Y_a)$ is L then $\mu(W)$ is F*
- If $\mu(A)$ is L, $\mu(Y_a)$ is H then $\mu(W)$ is F*
- If $\mu(A)$ is H, $\mu(Y_a)$ is L then $\mu(W)$ is F*
- If $\mu(A)$ is H, $\mu(Y_a)$ is H then $\mu(W)$ is T*
- If $\mu(G)$ is L, $\mu(Y_g)$ is L then $\mu(W)$ is F*
- If $\mu(G)$ is L, $\mu(Y_g)$ is H then $\mu(W)$ is F*
- If $\mu(G)$ is H, $\mu(Y_g)$ is L then $\mu(W)$ is F*
- If $\mu(G)$ is H, $\mu(Y_g)$ is H then $\mu(W)$ is T*
- If $\mu(C)$ is L, $\mu(Y_c)$ is L then $\mu(W)$ is F*
- If $\mu(C)$ is L, $\mu(Y_c)$ is H then $\mu(W)$ is F*
- If $\mu(C)$ is H, $\mu(Y_c)$ is L then $\mu(W)$ is F*
- If $\mu(C)$ is H, $\mu(Y_c)$ is H then $\mu(W)$ is T*
- If $\mu(T)$ is L, $\mu(Y_t)$ is L then $\mu(W)$ is F*
- If $\mu(T)$ is L, $\mu(Y_t)$ is H then $\mu(W)$ is F*
- If $\mu(T)$ is H, $\mu(Y_t)$ is L then $\mu(W)$ is F*
- If $\mu(T)$ is H, $\mu(Y_t)$ is H then $\mu(W)$ is T*

3. 실험 및 결과 분석

3.1 실험 환경

본 연구의 실험 환경은 삼성 Sens 노트북 x10

(M) 1.3GHz CPU와 512M RAM이 장착된 PC상에서 VC++ 6.0으로 구현하였다.

테스트 데이터를 얻기 위해서 완성된 단백질 지놈인 'Synechocystis PCC6803'을 실험 데이터로 적용하였다. 이 단백질의 염기 길이는 약 350만 개이므로 실험을 위해서 각각 1만개, 10만개를 추출하여 600~700개의 크기를 가진 단편을 생성하였으며, 이 단편으로 임의의 mutation을 유발하였다. 제안된 DNA 염기 서열 결합 결과 화면은 그림 5와 같으며 최종 DNA 염기 서열 결과는 그림 6과 같다.

3.2 실험 결과 분석

각 1만개, 10만개를 추출하여 mutation을 유발하여서 본 연구의 방법을 실험한 결과, 모두 original sequence를 구성하였으며, 실행 시간은 단편의 수에 비례하는 것을 알 수 있었다.

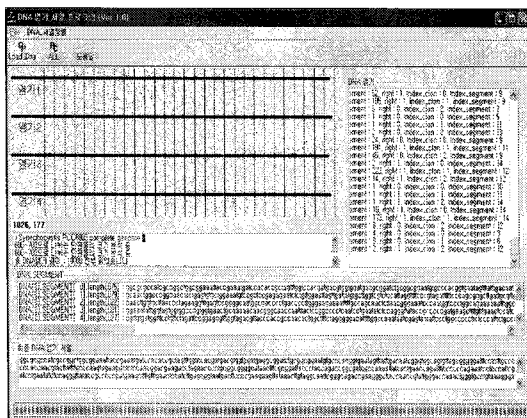


그림 5. 서열 결합 프로그램 인터페이스



그림 6. 최종 DNA 염기

표 1. 염기 추출 수에 따른 결합 시간

	FAP	Synechocystis PCC6803
10,000	48 sec	53 sec
100,000	487 sec	504 sec

표 1에서, FAP는 일치율만 사용함으로써 결합이 전부 완성 되지 않은 경우가 발생하였으나, 본 연구의 방법은 퍼지 추론 기법을 적용하여 가장 소속 정도가 높은 비교 단편을 결합하기 때문에 결합 실패가 발생하지 않았다. 서열을 결합하는데 가장 많은 시간이 소요되는 부분은 단편 쌍의 일치율을 측정하는 부분이다. 표 1에서와 같이 FAP보다 제안된 방법은 일치율을 측정하는 시간을 감소하기 위해서 compute agreement 알고리즘을 적용하여 계산하였기 때문에 DNA 단편 결합에 소요되는 시간이 기존의 FAP에 비해 감소하였다.

4. 결 론

기존의 서열 단편을 결합하는 방법들은 DNA 염기 단편 쌍들 간의 일치율 정보만으로 결합하였기 때문에 결합이 실패 하는 경우가 있었다. 본 연구에서는 이를 개선하기 위해 퍼지 추론 기법을 이용하여 일치율이 적더라도 이전의 빈도수와 각 A,G,C,T 염기의 소속도를 계산하여 결합 여부를 판단하는 알고리즘을 적용하였다.

본 연구에서 제안된 DNA 서열 결합하는 방법은 매우 긴 DNA의 염기서열을 자동 서열 분석기로 한번에 분석 가능한 약 700개의 단편들을 한 주형으로 만들어 PCR 방법으로 클론을 3개 생성 후, 600~700개의 길이로 단편화하여 기준 주형과 비교하여 일치율을 계산하였다. 이때 일치율의 계산 시간을 위하여 Compute Agreement 알고리즘을

이용하여 일치율을 계산하는 시간을 단축시켰다. 계산된 단편 쌍들의 중첩정도를 기준으로 주형마다 2개의 결합 후보 단편을 추출하여 추출된 각 단편들의 일치율과 각 A,G,C,T의 소속도 및 각 A,G,C,T의 이전 빈도수를 퍼지 추론 규칙에 적용하여 결합 여부를 판단하였다. 본 연구에서는 결정된 최적의 비교 단편을 결합하고, 더 이상 단편이 없을 때까지 반복하여 서열 결합을 완성하였다.

본 연구의 테스트 데이터를 얻기 위해서 완성된 단백질 지놈인 'Synechocystis PCC6803'을 실험 데이터로 적용하였다. 이 단백질의 염기 길이는 약 350만개 이므로 실험을 위해서 각각 1만개, 10만개를 추출하여 600~700개의 크기를 가진 단편을 생성하였으며, 이 단편을 임의의 mutation을 유발 시켜서 실험한 결과, FAP 프로그램보다 속도가 줄어들었으며, contig 구성 프로그램의 단점인 결합 실패가 발생하지 않았다. 최종 결합된 단편은 모두 original sequence를 구성함으로써 이전의 방법보다 개선됨을 확인하였다.

향후 방향은 Needleman-Wunsch의 DP 기반 알고리즘에서 행렬 생성 단계에서 발생하는 불필요한 정렬 계산을 제거하여 전체 수행 시간을 단축하고 지능적으로 깎 비용을 동적으로 조정하는 방법에 대해 연구할 것이다.

참 고 문 헌

[1] Staden, "A new computer method for the storage and manipulation of DNA gel reading data," Nucl. Acids, Res. 8, pp. 3673-3694, 1980.
 [2] Hannu, P., H. Soderlund and E. Ukkonen, "SEQAID: a DNA sequence addembling program based on a mathmedical model," Nucl. Acids, Res. 12, pp. 307-321, 1984.

[3] Xiaociu, H, "A Contig Assembly Program Based on sensitive Detection of Fragment Overlaps," Genomics, Res. 14, pp. 18-25, 1992.
 [4] 이병욱, 박기정, 박완, 박용하, "DNA 염기 서열의 단편 조립 프로그램 개발," Kor. J. Appl. Microbiol. Biotechnol. 제25권, 6호, pp. 560-565, 1997.
 [5] George J. K. and Bo Y., Fuzzy Sets and Fuzzy Logic Theory and Applications, Prentice Hall PTR, 1995.
 [6] 김광백, 박현정, "퍼지 추론 기법을 이용한 DNA 염기 서열 단편 결합," 한국해양정보통신학회 논문지, 10권, 12호, pp. 2329-2334, 2006.
 [7] Sanger, F., Nickden, S., and Coulson, A.R. "DNA Sequencing with chain terminator inhibitors," Proc. Natal. Acad. Sci. USA 74, pp. 5463-5467, 1977.



김 광 백

- 1999년 부산대학교 전자계산학과(이학박사)
- 1997년~현재 신라대학교 컴퓨터공학과 부교수
- 1999년~2000년 Biomedical Fuzzy Systems Association Associate Editors (Japan)
- 2005년~현재 한국멀티미디어학회 이사 및 논문지 편집분과위원장
- 2005년~현재 한국해양정보통신학회 이사 및 논문지 편집 부위원장
- 관심분야 : Neural Networks, Image Processing, Fuzzy Logic, Medical Imaging and Biomedical System, Support Vector Machines



우 영 운

- 1989년 2월 연세대학교 전자공학과(공학사)
 - 1991년 8월 연세대학교 본대학원 전자공학과(공학석사)
 - 1997년 8월 연세대학교 본대학원 전자공학과(공학박사)
 - 1997년 9월 ~ 현재 동의대학교 멀티미디어공학과 교수
 - 관심분야 : 인공지능, 패턴인식, 퍼지이론, 의료정보
-
-