제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

정규화된 지역 정렬 알고리즘을 적용한 다중 지역 정렬 알고리즘

장석봉* 이계성**
* ** 단국대학교 전자컴퓨터학부
e-mail:hollo98@korea.com

An Algorithm for multiple local alignment with Normalized Local Alignment Algorithm

Suk-Bong Jang* Gye-Sung Lee**
* ** Dept of Computer Science and Electronics, Dan-Kook University

요 약

두 서열을 비교하여 유사성(similarity)이나 상동성(homology)를 찾기 위한 서열 정렬 방법 중에서 지역 정렬에 많이 사용되는 Smith-Waterman 알고리즘의 제한점인 Mosaic effect와 Shadow effect를 극복하기 위한 효율적인 방법을 살펴보고, 하나의 최대값이 아닌 다수개의 최대값을 찾아 다수개를 정렬함으로써 서열내에 존재 할 수 있는 다수개의 지역 정렬을 찾고 Normalized sequence alignment 알고리즘을 이용하여 서열 정렬된 결과들의 우선 순위를 매겨본다.

1. 서론

생물 정보학에 있어서 매우 중요한 분야의 하나인 서열정렬이란 핵산(DNA)나 단백질(protein)의 서열을 적절히 배열시켜 두 서열간의 유사성(similarity)이나 상동성(homology)을 판단할 수 있는 방법이다. 상동성은 두 가지로 나뉘어 지는데 전역 정렬은 같은 종류의 핵산이나 단백질 서열을 비교하여 최대의 상동성이 나타나도록 하는 방법이고 대표적인 알고리즘으로 Needleman-Wunsch 알고리즘이 있다. 전역 정렬의 경우는 두 서열의 시작에서 끝까지를 비교하여 가장 높은 점수를 가지도록 정렬하는 것이다.

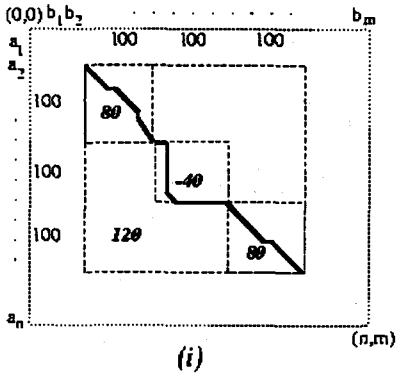
상동성 판단을 위한 다른 방법인 지역 정렬은 전역 정렬과는 다르게 전체를 비교하는 것이 아닌 두 서열간의 일치하는 부분을 찾아 정렬시킨다. 이 경우는 두 서열이 공통된 도메인을 공유하거나, 다변화된 서열을 비교, 특히 다른 종으로 구분되지만 진화적으로 공통된 원점을 공유하는 경우에 유용하게 사용되어진다. 지역 정렬에 사용되는 대표적인 알고리즘으로 Smith - Waterman 알고리즘이 있다.

2. Mosaic effect 와 Shadow effect

지역 정렬 문제는 길이가 서로 같거나 다른 두 서열 $n \geq m$ 에서 유사한 부분을 찾는 것이다. 그러나 지역 정렬의 일반적인 개념에서는 찾아진 서열 즉 지열 정렬이 된 서열의 길이를 고려하지 않음으로 인해 발생할 수 있는 문제점들이 있다 [1]. 예를 들어 찾아진 정렬의 score가 1000이고 길이가 10000인 지역 정렬과 score가 998이고 길이가 1000인 지역 정렬이 있을 때 고전적인 Smith-Waterman 알고리즘은 생물학적으로 덜 중요할지라도 점수값이 큰 정렬을 찾게 된다 [2]. 이러한 문제점이 지역 정렬에 일반적으로 사용되는 Smith-Waterman 알고리즘은 최고의 점수를 가지는 정렬을 찾아주기는 하지만, 최고의 유사성을 가지는 서열을 찾는데는 제한점으로 나타나게 된다.

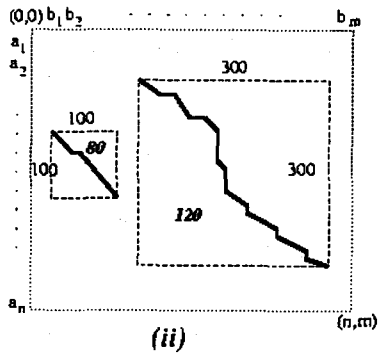
두 서열을 비교 할 경우 Smith-Waterman 알고리즘은 최고의 점수를 가지는 서열을 찾아주기는 하지

만 찾아진 서열의 길이를 고려하지 않았기 때문에 Mosaic effect 와 Shadow effect 가 나타나게 된다. Mosaic effect는 Smith-Waterman 알고리즘이 찾은 지역 정렬에서 매우 하찮은 영역이 높은 similarity score를 가지는 두 개의 영역 사이에 끼워져 있을 때 <그림 1>과 같이 관찰되어진다 [2].

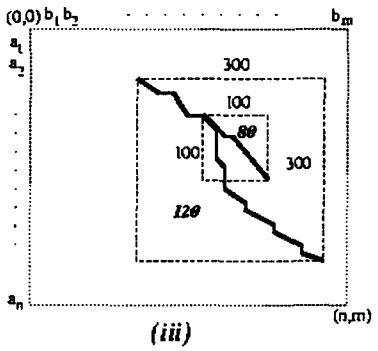


<그림 1> Mosaic effect

또 다른 제한점으로 나타나게 되는 Shadow effect 는 위에서 말한 정렬의 길이와 score의 경우처럼 생물학적으로 부적당할지라도 단지 약간 더 높은 score를 가지는 이유만으로 길이는 짧지만 생물학적으로 중요한 정렬이 탐지되지 않을 때 관찰되어진다. 이 경우 <그림 2>에서처럼 non-overlapping되어 지거나 <그림 3>과 같이 overlapping되어질 수 있다.



<그림 2> Shadow effect
(non-overlapping alignments)



<그림 3> Shadow effect
(overlapping alignments)

이렇게 <그림 2>와 <그림 3>의 경우처럼 overlapping 이거나 non-overlapping일 경우는 계산되어진 점수행렬 내에서 하나의 최대값을 찾아 지역 정렬을 찾게되는 Smith-Waterman 알고리즘의 제한점으로 인해 제외되어지는 결과를 가져오므로 개선된 알고리즘이 요구된다.

3. Smith-Waterman 알고리즘의 제한점과 개선책

Smith-Waterman 알고리즘은 computational 분자생물학에서 매우 중요하게 사용되는 기법의 하나로 잘 보존된 부분을 찾고 그렇지 못한 부분은 제거하도록 설계되어졌다. 정렬에 사용되는 두 서열, X와 Y가 아래와 같이 정의된다.

$X = x_1 x_2 \ldots x_n$와 $Y = y_1 y_2 \ldots y_m$ ($n \geq m$)

여기서 지역 정렬이란 서열 X 와 Y의 부분 서열, 즉, I 와 J의 쌍을 포함하는 정렬을 의미하고, 지역 정렬 점수인 $S_{i,j}$ 가 할당되는데, 이 값을 최대화하는 Dynamic Programming 공식은 다음과 같이 정의된다:

$$S(i,j) = \max\{\ 0,\ S(i-1,j)-d,\ S(i-1,j-1)+s(x_i,y_j),\ S(i,j-1)-d\ \}$$

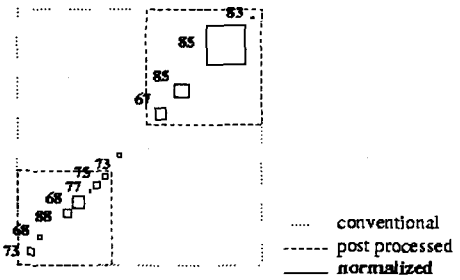
이 알고리즘을 이용한 지역 정렬은 score matrix상에서 가장 큰 값을 갖는 점으로부터 역방향 으로 추적하면서 정렬을 하게 된다.

결과적으로 Smith-Waterman 알고리즘은 유사성의 척도만을 가지고 두 서열간의 정렬을 통해 지역 정렬을 찾게 되므로 서열 정렬에서 비유사도(non-similarity)가 높은 시작 부분과 끝 부분은 효율적으로 제거할 수 있으나 그 부분이 서열의 사이

에 있을 경우 이를 하나로 병합시키는 결과를 가져온다. 이러한 문제점으로 인해 앞에서 살펴본 Mosaic effect 나 Shadow effect가 나타나게 되고 이를 개선하기 위한 많은 연구들이 있었다. 그중 Zhang 이 제시한 방법으로 지역 정렬을 완전히 구성한 뒤 부분 서열 정렬이 이뤄지도록 하여 <그림 1>에서 보이는 -40의 score를 가지는 부분을 가지지 않도록 하는 방법이 이다 [3]. 그러나 이 방법은 지역 정렬내에 포함이 되지 않으면 의미 있는 부분 정렬이 있을 경우 찾지 못한다는 것이다. 또 다른 해결 방법으로 Arslan and Eğecioğlu이 제시한 것으로 $|I|+|J| \geq t$ 인 모든 부분 서열 I 와 J 사이에서 $(S_{i,j}) / (|I|+|J|)$ 를 최대화하는 I 와 J를 발견하는 Normalized sequence alignment를 제안했다. 여기에서 $S_{i,j}$는 score, t는 I 와 J 의 minimal length를 위한 임계값이고 아래와 같이 표현된다:

$$NAL \approx \max\{s(I,J)/(|I|+|J|) | I \subseteq X, J \subseteq Y, |I|+|J| \geq t\}$$

이 방법을 적용하면 <그림 1><그림 2><그림 3>에서 볼 수 있는 것처럼 임계값 t의 길이를 200으로 한 경우 얻을 수 있는 값은 80/200 = 0.4, 600일 때 120/600 = 0.2를 얻게 되어 길이가 짧은 서열로 나누었을 때 더 나은 점수를 얻을 수 있는 것을 알 수 있다 [2]. 이 실험에서 기존의 S-W 알고리즘과 Zhang의 post-processed local alignment그리고 normalized를 비교한 것을 보면 아래의 <그림 4>와 같다 [1].



<그림 4> 비교 실험결과

4. 다중 지역 정렬을 위한 알고리즘

알고리즘 : score matrix 내에 하나가 아닌 다중의 지역 정렬을 찾기 위한 방법으로 Dynamic Programming Algorithm 에 의해 계산되는 과정에서 이전의 최대값과 현재 위치에서의 값을 비교하여 값을 갱신하거나 유지 해나간다. 이때 값들의 위치를 고려해야 하는데 그 이유는 다수개의 최대값을 찾기 위해 계산되어지는 값과 그 값의 위치를 계속해서 유지하게 된다. 계산이 계속 수행되면서 현재 유지된 값과 다음에 오게될 값의 score matrix 상의 거리가 충분히 멀 경우 그 값은 또 다른 하나의 최대값으로 구분을 하게 되고 그렇지 않을 경우는 Dynamic Programming Algorithm 의 특성으로 나타나게 되는 이전의 최대값과 유사한 값으로 판단한다.

이 연구에서 다수개의 최대값을 찾기 위한 방법을 간단히 정리하면 다음과 같다.

입력은 다음과 같이 두 서열이 주어진다.

$X = x_1 x_2 \ldots x_n$와 $Y = y_1 y_2 \ldots y_m$ ($n \geq m$)

그리고 다수개의 최고값을 위해 정의되는 Maxscore = [K]에서 K는 다중 지역 정렬의 수가되고, K개의 다수개의 최대값을 유지한다는 의미로 값을 제한하는 이유는 다수개의 값을 찾는 것이 목적이지만 너무 많은 값을 찾는 것은 별 의미 없는 자잘한 조각까지 찾게되는 결과를 가져오게 되므로 두 서열 X와 Y의 길이를 고려하여 그 수를 제한한다. K의 값은 계속해서 sort 되어지면서 유지하게 된다.

1. score matrix의 최 좌측열은 0으로 채워지고 Dynamic Programming Algorithm 에 의해 계산을 수행한다.
   (이때 다수개의 최대값을 취하기 위해 현재의 값과 이전의 값을 비교하여 값과 위치를 수렴하게 된다.)
2. 최초에 만나게 되는 $0 < S_{i,j}$ 값과 그 위치를 최고값으로 저장한다.
3. 계산이 수행중인 $S_{i,j}$ 가 저장된 최고값 보다 크면 값과 위치를 저장한다. (단 이 값은 유지된 값의 위치와 현재의 위치(두 값의 distance)를 고려하여 갱신되거나 유지되어진다.) 그렇지 않을 경우 다음 계산을 수행한다.
4. K개의 최대값 위치를 시작점으로 지역 정렬을 한다.
5. NLA Alogrithm score를 적용하여 서열정렬 값을 최대로 해주는 K개의 서열을 찾는다.
6. 위의 점수를 이용하여 서열의 생물학적 우선순위를 매긴다.

5. 실험 결과

주어진 2개의 단백질 서열에서 다수개의 최대값을

찾고 이 값을 이용하여 지역 정렬을 찾는 시험을 해 보기로 한다. 실험에는 BLOSUM 50 대체 행렬을 사용하여 실험하였다. 실험에 사용된 두 서열은 다음과 같이 정의되었다.

서열 1 : HEAGAWGPHEE

서열 2 : PAWTHEAE

두 서열에서 기본적인 Smith-Waterman 알고리즘은 정렬 점수가 0이 되는 지점에서 정렬을 멈추게 되어 서열내에 포함된 조각부분(gap과 mismatch)을 포함하여 <그림 5>와 같은 하나의 긴 서열을 출력하게 된다. 이렇게 지역 정렬된 경우는 최종적으로 27점을 가지는 정렬이 된다.

서열 1	A	W	G	P	H	E
서열 2	A	W	-	T	H	E

<그림 5> 일반적 지역 정렬 결과

아래의 <표 1>은 matrix 는 두 서열의 점수 계산 결과이다.

	x	0	1	2	3	4	5	6	7	8	9	10	11
y			H	E	A	G	A	W	G	P	H	E	E
0		0	0	0	0	0	0	0	0	0	0	0	0
1	P	0	0	0	0	0	0	0	0	10	2	0	0
2	A	0	0	0	5	0	5	0	0	2	8	1	0
3	W	0	0	0	0	2	0	20	12	4	0	5	0
4	T	0	0	0	0	0	2	12	18	11	3	0	4
5	H	0	10	2	0	0	0	4	10	16	21	13	5
6	E	0	2	16	8	0	0	0	2	9	16	27	19
7	A	0	0	8	21	13	5	0	0	1	8	19	26
8	E	0	0	6	13	18	12	4	0	0	1	14	25

<표 1> 두 서열의 Score Matrix

위의 <그림 5>의 고전적 지역 정렬 결과와 <표 1>의 score matrix를 보면 가장 큰값은 27이고 이 값을 기준으로 정렬을 고전적인 방법에서는 gap부분과 mismatch 부분을 제거하더라도 다른 의미 있는 부분을 찾을 수 없다. 하지만 위에서 소개한 방법을 이용하여 최고값을 유지해 나가게 되면 먼저 21(음영으로 보이는)점을 유지하게 되고 다음에 오게되는 의미 있는 값($0<S_{i,j}$)들과의 distance를 고려하여 20점을 가지는 지역과 27점을 가지는 3개의 시작점을 얻을 수 있고 이 위치를 기준으로 다중 지역 정렬을 찾을 수 있다.

x	5	6	v	2	3	x	8	9	v	4	5
	A	W		A	W		H	E		H	E
20 점 (5점)						16 점 (4점)					
x	4	5	6	v	1	2	3				
	H	E	A		H	E	A				
21 점 (3.5점)											

<표 2> test 서열의 다중 지역 정렬 결과

<표 2>는 개선된 방법을 이용하여 지역 정렬을 하였을 때의 정렬 결과와 지역 정렬들이 가지는 점수를 보여준다. ( )안의 score는 score/|I|+|J| 의 점수. 즉. 정렬의 길이를 고려한 점수이다. 우리는 이 점수에서 볼 수 있듯이 가장 큰 정렬 점수를 가지는 지역 정렬로 HEA → AW → HE 의 순서로 순위를 찾을 수 있지만, 지역 정렬의 길이를 고려한 normalized local alignment 기법을 적용한 점수를 비교해보면 우선 순위는 AW → HE → HEA 의 순서로 달라지게 된다.

5. 논의 및 결론.

본 연구에서는 Smith-Waterman 알고리즘의 제한점으로 나타나게되는 Mosaic effect와 Shadow effect에 대해 살펴보고 이를 극복하기 위한 방법 중 normalized sequence alignment score[2]를 적용하여 다중 지역 정렬을 시도하였다. 기존의 지역 정렬을 위한 시도들이 포함하지 않았던 부분을 포함할 수 있도록 고안하였고 normalized 기법을 적용해 서열의 길이를 고려함으로써 좀더 개선된 알고리즘으로 발전을 시켰다. 그러나 간단한 test 서열을 이용한 실험이기에 좀더 체계적인 분석과 효율성의 검증이 요구된다.

참고문헌

[1] A. N. Arslan, Ö. Eğecioğlu, and P. A. Pevzner "A new approach to sequence comparison : Normalized sequence alignment" Bioinformatics, 17(4):327-337, 2001.

[2] A. N. Arslan, Ö. Eğecioğlu "Algorithms for Local Alignments with Length Constraints" , Proc. 5th Latin American Theoretical Informatics Symposium (LATIN 2002), Lecture Notes in Computer Science 2286, pp. 38-51, Cancun, Mexico, April 2002.

[3] Z. Zhang, P. Berman, T. Wiehe, and W. Miller "Post-processing long pairwise alignments" Bioinformatics, 15, 1012-1019, 1999.