Local chaining 알고리즘의 단점 및 개선 방법

이선호^{O} 박근수
서울대학교 전기,컴퓨터공학부
{shlee^{O}, kpark}@theory.snu.ac.kr

Improving Weaknesses of Local Chaining Algorithms

Sunho Lee^{O} Kunsoo Park
School of Computer Science & Engineering, Seoul National University

요 약

Chaining 알고리즘은 주어진 match 정보로부터 좋은 match 조합을 찾아내는 일종의 alignment 알고리즘으로 유전체 서열을 비교하는데 다양하게 응용되고 있다. 특히 서열 전체를 비교하는 대신 부분 서열을 비교할 때 사용할 수 있는 local chaining 알고리즘이 제안되었는데 본 논문은 이 기본적인 알고리즘이 Smith-Waterman 알고리즘과 유사하며 따라서 비슷한 단점을 가지고 있음을 지적한다. 그리고 이를 해결하기 위해 X-drop과 정규화된 점수를 고려하는 두 가지 기법을 적용하고 실험을 통해 개선 효과를 보인다.

1 서 론

1.1 Chaining 기법

유전체 서열을 비교하는 프로그램에서 먼저 비슷한 부분 match들을 찾은 다음 발견한 match들을 연결해 답을 내놓는 접근 방법이 널리 쓰이고 있다[2, 4, 5, 10]. 비슷한 부분 match들을 연결해 좋은 조합을 찾는 문제는 '좋은 조합' 기준에 따라 달라지지만 일반적으로 많이 쓰이는 chaining 문제는 다음과 같다. 입력으로 두 서열 사이의 비슷한 부분 match들의 정보를 받아 출력으로 match들 사이의 상대적인 순서가 보존되고 서로 겹치지 않는 조합 중 가장 점수 총합이 높은 조합을 찾는다.

1.2 Chaining 문제 정의

각 match들을 두 서열을 x축 y축으로 한 2D 평면상에 사각형으로 나타낸다. 하나의 match 사각형 f는 시작점 $f.beg$와 끝점 $f.end$를 가지고 x축 서열의 $f.beg.x \sim f.end.x$ 영역과 y축 서열의 $f.beg.y \sim f.end.y$ 영역이 서로 비슷하다는 정보를 나타낸다. 그리고 각 match f의 비슷함 정도에 따라 점수 $f.weight$를 갖는다.

두 match f, f' 사이에는 순서 관계와 연결 비용을 갖는다. 순서 관계는 f가 f'의 왼쪽 아래에 있을 때 $f < f'$ 라고 정의하고 $f < f'$ 이면 f, f'은 서로 겹치지 않고 x축, y축 서열 상에서 상대적인 순서가 보존된다. 연결 비용은 $gap(f, f')$로 정의하고 여러 가지 방법이 있지만 보통은 다음처럼 대각선 길이와 직선 길이 합을 사용한다[1, 9].

$$gap(f, f') = r(dx - dy) + edy = rdx + (e-r)\,dy \text{ if } dx > dy$$
$$r(dy - dx) + edx = rdy + (e-r)\,dx \text{ if } dy \geq dx$$
(r : insert/delete 비용, e : substitution 비용)

chain은 순서 관계에 따라 match들을 늘어놓은 순열로 정의하고,

$$\text{a chain} = \{ f_1, f_2, \ldots f_n : f_i < f_{i+1} \}$$

이 순열의 부분 순열을 subchain으로 정의한다.

$$\text{a subchain of } \{f_i\}_{i=1..n} = \{ f_{j+1}, f_{j+2}, \ldots, f_{j+N} : j \geq 0 \text{ and } N \leq n-j \}$$

즉 subchain은 전체 chain이 대표하는 두 서열 영역의 일부 영역을 대표하는 chain이 된다.

각 chain은 match들의 비슷함 점수 $f_i.weight$ 합과 연결 비용 $gap(f_i, f_{i+1})$차로 총점을 갖는다.

$$f.score(C) = \Sigma_i f_i.weight - gap(f_i, f_{i+1})$$

주어진 match 집합에서 $s(0,0)$에서 시작해 $t(n,m)$에서 끝나는 chain C중에서 가능한 총점 $score(C)$가 가장 높은 chain을 찾는 문제를 global chaining 문제로 정의한다.

1.3 Chaining 알고리즘

점수의 합이 가장 높은 chain을 구하는 간단한 방법은 subchain이 전체 chain에 독립인 점을 이용한 dynamic programming 기법이다[1, 7, 9].

각 match f에 대해 $f.score$를 f에서 끝나는 가장 좋은 chain의 총점으로 정의하면,

$$f'.score = \max\{ score(C) : C \text{ is a chain ending with } f' \}$$

$f.score$는 다음의 점화식을 이용해 구할 수 있다.

$$f'.score = f'.weight + \max\{ f.score - gap(f, f') : f < f' \}$$

이러한 global chaining 알고리즘으로 구한 chain은 일종의 global alignment로 서열 전체 영역 사이의 비슷함을 보여준다. 따라서 두 서열이 전체적으로 매우 비슷한 경우에 의미 있는 결

과를 얻을 수 있지만 두 서열이 부분적으로만 비슷한 경우 global alignment는 의미가 없을 수 있다.

1.4 Local chaining 알고리즘

유전체 서열을 비교하는 목적에 따라 전체 영역 대신 부분적인 영역 사이의 비슷함을 보여주는 chain이 필요하다. 어떤 주어진 서열을 포함하는 가장 비슷한 영역을 찾거나 두 서열 사이에 어느 정도 점수가 좋은 부분적인 chain들을 구할 필요도 있다[2, 5]. Global chaining 문제를 변형해 시작과 끝 위치 제약을 없애고 항상 chain의 점수가 양수인 chain중 가장 좋은 chain을 찾는 문제로 바꾸면 local chaining 문제가 된다. 즉 local chain의 시작과 끝 위치는 어느 match도 가능하고 local chain의 총점은 특정 값 0을 넘어야 한다.

Local chain은 두 서열 사이에 local alignment를 구하는 Smith-Waterman 알고리즘과 비슷한 방법으로 구할 수 있다. Global chaining 알고리즘을 약간 변경해 점수가 0이상인 경우에만 chain을 계속 연결하고 비슷한 부분들이 멀리 떨어져 있어서 연결 비용이 커지면 연결을 포기하고 새로운 chain을 시작한다[2].

$$f'.score = f'.weight + \max\{\, 0, f.score - gap(f, f') : f < f' \,\}$$

이러한 알고리즘은 가장 좋은 local chain을 찾는 방법이지만 발견한 chain에 속하는 match들을 제외하고 다시 알고리즘을 적용하면 점수가 높은 순서대로 특정 값 이상의 chain을 모두 구할 수 있다.

2 단점과 개선 방법

2.1 Local chaining 알고리즘의 단점

Smith-Waterman 알고리즘과 흡사한 local chaining 알고리즘은 그 동안 지적되었던 Smith-Waterman 알고리즘의 단점들[3, 8]과 비슷한 단점을 갖는다.

구한 local chain내에 멀리 떨어져 있는 상관없는 match들이 끼어 들 수 있다. 이를 모자이크 효과라고 하며 앞부분이나 뒷부분의 좋은 subchain 때문에 중간에 점수가 별로 좋지 않은 subchain이 끼어 들 수 있다.

그리고 길이가 긴 chain 근처에 짧지만 생물학적으로 의미 있는 chain을 찾지 못 할 수 있다. 이를 그림자 효과라고 하며 대체로 길이가 긴 chain이 점수가 높아지는 경향이 있어서 짧지만 좋은 chain이 긴 chain에 가려질 수 있다.

이런 문제를 해결하기 위해 Smith-Waterman 알고리즘의 문제를 해결하는데 사용한 방법들을 적용해 볼 수 있다. 이 논문에서는 다음의 두 가지 방법을 chaining에 적용해 보았다. 첫번째는 local alignment 점수가 특정 값보다 작은 영역(X-drop)을 포함시키지 않는 방법[11]이고 두번째는 단순한 local alignment 점수 대신 길이를 고려해 정규화된 점수를 고려하는 방법[3. 6]이다.

2.2 X-drop chaining

비슷한 부분 match를 연결해 chain을 만들 때 멀리 떨어져 있어서 연결 비용이 큰 match들은 chain에 연결시키지 않는다. 이것은 일종의 X-drop 접근 방법으로 chain 점수가 특정 값보다 작은 영역이 항상 match와 match 사이의 연결 부분이 되며 다음과 같이 점화식을 변형시키면 구할 수 있다.

$$f'.score = f'.weight + \max\{\, 0, f.score - gap'(f, f') : f < f' \,\}$$
$$gap'(f, f') = \begin{cases} gap(f, f') & \text{if } gap(f, f') < T \\ \infty & \text{if } gap(f, f') \geq T \end{cases}$$

2.3 Normalized chaining

Local chain 점수를 그 chain이 대표하는 부분 서열 길이의 합으로 나눈 새로운 정규화된 점수를 도입해서 특정 값 이상의 정규화된 점수를 갖는 local chain을 찾는다.

$$f..nscore = f.score / (f.length + L)$$

match f'에서 시작해 f로 끝나는 chain의 길이 $f.length$는 chain이 대표하는 x축, y축 서열 길이의 합으로 정의 된다. 상수 L은 너무 짧은 chain이 선택되지 않도록 하는 효과를 갖는다.

$$f.length = f.end.x + f.end.y - (f'.beg.x + f'.beg.y)$$
(f'은 chain의 시작 match)

정규화된 chain을 구하는 점화식은 다음과 같다.

$$f^* = argmax_f \frac{f'.weight + f.score - gap(f, f')}{f.length + f'.end.x + f'.end.y - (f.end.x + f.end.y) + L}$$

($f < f'$ and $f'.nscore > NT$, 이런 f가 없으면 f^* = NULL)

$$f'.score = f'.weight + f^*.score - gap(f^*, f')$$
$$f'.length = f^*.length + f'.end.x + f'.end.y - f^*.end.x - f^*.end.y$$

그러나 이 식은 앞의 점화식들과 달리 항상 최적의 정규화된 점수 $f.nscore$를 유지하는 것을 보장하지 못하는 근사적인 방법이다.

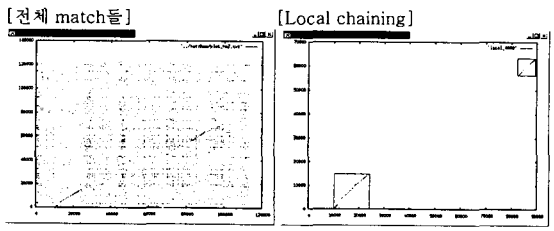
3 실험 결과

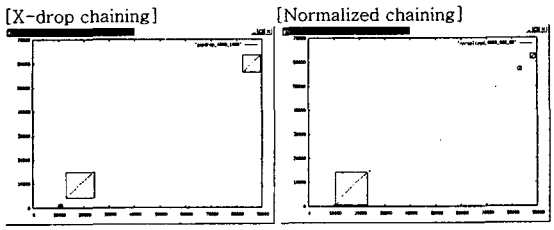
실제 생물 데이터에 대해 제안한 두 가지 처리 방식이 효과가 있는지 확인해 보았다. 사람과 쥐의 DNA서열 데이터[3, 5]를 가지고 먼저 12bp 이상의 exact-match들을 찾아 이것을 match 데이터로 했다. match를 찾는 방법은 여러 가지가 있지만 여기서는 간단하게 exact-match를 사용했다. 각 match의 점수는 exact-match 길이의 10배 값을 사용하고 연결 비용은 대각선 길이와 직선 길이의 합을 그대로 사용했다.($e = r = 1$)

그리고 특정 값 이상의 local chain을 모두 구하기 위해 위의 알고리즘을 반복 적용했다. 간단한 local chaining은 총점 4000점 이상의 모든 chain, X-drop chaining은 연결 비용 1000미만이고 총점 4000점 이상, Normalized chaining은 연결 허용(NT)은 0.05이상, L = 4000, 정규화된 총점은 0.2이상의 모든 chain을 구했다.

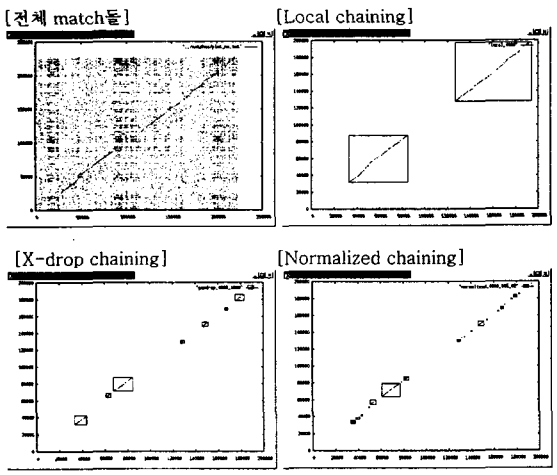
실험1 – Genbank accession no. AF030876/ AF121351

[전체 match들] [Local chaining]





실험2 – Genbank accession no. U47924/ AC002397



위의 두 실험 결과를 보면 간단한 local chain에 비해 좀 더 가까운 match들끼리 묶어서 짧지만 좋은 chain을 만들 수 있음을 알 수 있다. 실험 결과의 총점 평균을 비교해 보면 연결 비용에 제한을 주거나 정규화된 점수를 사용한 쪽이 총점이 낮아지는 대신 match 개수 당 총점은 증가하는 것을 볼 수 있다. 즉 짧지만 match들이 가까이 있어서 총점이 높은 chain을 찾을 수 있다.

* 총점 : chain의 총점, 개수 : chain을 이루는 match의 수

		총점/개수	총점평균	개수평균	최대총점	최대개수
실	기본	167.90	18385.0	109.5	27450	145
험	X-drop	179.93	13074.7	72.7	25500	120
1	정규화	192.07	6818.5	35.5	26772	134
실	기본	56.62	25225.5	445.5	27403	475
험	X-drop	114.58	8564.6	74.8	15011	122
2	정규화	128.19	4271.1	33.3	13171	83

실험 결과는 서열 데이터와 인자에 따라 달라질 수 있으며 실제 생물학적 응용에서는 인자를 조절해 좀 더 의미 있는 match들의 조합을 찾는 기능을 제공할 수 있다.

4 결론 및 향후 과제

Local chaining 알고리즘이 Smith-Waterman 알고리즘과 비슷한 단점을 가지고 있으며 같은 접근 방법으로 해결할 수 있음을 알아보았다. 그러나 기본적인 local chaining은 효율적인 풀이 방법이 많이 알려져 있어서 일반적인 $O(n^2)$ dynamic programming 기법 보다 빠르게 풀 수 있는 데 비해[1, 2, 9], 제안한 방법에 대해서는 효율적인 알고리즘이 알려져 있지 않다. 생물학 응용에 적합하면서 보다 효율적인 알고리즘의 개발이 앞으로의 과제가 될 것이다.

5 참고 문헌

[1] M. I. Abouelhoda and E. Ohlebusch. Multiple genome alignment: Chaining algorithms revisited. In Proceedings of he 14th Annual Symposium on Combinatorial Pattern Matching, LNCS 2676, 1-16, 2003.

[2] M. I. Abouelhoda and E. Ohlebusch. A local chaining algorithm and its applications in comparative genomics, In Proceedings of the 3rd Workshop on Algorithms in Bioinformatics, LNBI 2816, 1-16, 2003.

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

[4] S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander. Human and mouse gene structure: Comparative analysis and application to exon prediction. Genome Research, 10, 950-958, 2001.

[5] A. L. Delcher, A. Phillippy, J. Carlton, and S. L. Salzberg. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research, 30(11), 2478-2483, 2002.

[6] N. Efraty and G. M. Landau, Sparse normalized local alignment, manuscript, 2004.

[7] D. Gusfield. Algorithms on Strings, Trees, and Sequences, Cambridge University Press, 1997.

[8] 김상태, 임계값 이상의 정규화된 국소적 정렬을 찾는 효율적인 알고리즘, 서울대학교 석사학위 논문, 2002.

[9] E. W. Myers and W. Miller. Chaining multiple-alignment fragments in sub-quadratic time. In Proceedings of the 6th ACM-SIAM Symposium on Discrete Algorithms, 38-47, 1995.

[10] S. Schwartz, Z. Zhang, K. A. Fraser, A. Smit, C. Riemer, J. Bouck, R. Gibson, R. Hardisson, and W. Miller. Pipmaker – A web server for aligning two genomic DNA sequences. Genome Research, 10, 577-586, 2000.

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