MSMP 알고리즘과 RIFLE 알고리즘의 구현 및 성능비교 평가

김동회° 원영상 고영웅 김진
한림대학교 정보통신공학부
{kdh°, wonys, yuko, jinkim}@hallym.ac.kr

Implementation and Performance Evaluation of Comparing MSMP with RIFLE Algorithm

Donghoi Kim° YoungSang Won YoungWoong Ko Jin Kim
Division of Information Engineering & Telecommunications, Hallym University

요 약

생물정보학에서 서열의 유사성을 예측하는 것은 가장 중요한 문제 중의 하나이다. 염기 서열의 유사성을 검색하는 유용한 검색도구들에는 BLAST와 FASTA 등이 있으며 이러한 도구들은 새로운 유기체에 대한 실제 염기 서열을 필요로 한다. 이 경우 서열을 얻기 위한 sequencing 작업이 필요로 하며 시간적인 면에 있어서 상당한 비용을 요구한다. 본 논문에서는 sequencing 작업을 하지 않고 간단한 실험에서 얻을 수 있는 부분적인 sequence 정보만을 대상으로 데이타베이스에서 검색을 할 수 있는 두 개의 RIFLE(Rapid Identification of Microorganisms by Fragment Length Evaluation), MSMP(Maximum Site Matching Problem) 알고리즘을 구현하고 실험을 통해 두 알고리즘을 비교 평가한다. 실험결과 RIFLE 알고리즘이 수행 속도면에서 빠른 반면 MSMP가 산출한 결과에 비해서 신뢰성이 떨어짐을 확인하였다.

1. 서론

생물정보학은 생물학 관련 데이터를 컴퓨터를 이용 정리, 분석, 이용하는 연구하는 학문이다. 그 중 서열을 검색, 분석하는 문제는 분자생물학의 여러 분야에서 상당히 중요한 문제에 해당한다. 서열 검색은 알려지지 않은 유기체를 sequencing작업을 통하여 염기서열을 알아내고 기존에 밝혀져 있는 염기서열과 가장 유사한 서열을 검색하는 것이다. 염기서열의 유사성을 통하여 그 구조와 기능을 유추하려는 것이 서열의 유사성 검색의 목적이라고 할 수 있다. 서열의 유사성을 통하여 검색을 해주는 유용한 검색 도구들에는 BLAST[1]와 FASTA[2]등이 있다. 이러한 도구들은 새로운 유기체에 대한 실제 염기서열을 필요로 한다. 알려지지 않은 새로운 유기체를 sequencing작업을 통해서 염기서열을 알아내고 염기서열 데이터베이스에서 가장 유사한 서열을 검색하는 것이다. 이때 서열을 검색하기에 앞서 행하는 sequencing 작업은 시간적인 면에 있어서 상당한 비용을 요구한다. 본 논문에서는 sequencing작업을 하지 않고도 염기서열 데이터베이스에서 검색을 할 수 있는 Restriction Pattern을 이용한 서열 검색 알고리즘인 RIFLE(Rapid Identification of Microorganisms by Fragment Length Evaluation)[3]과 MSMP(Maximum Site Matching Problem)[4]알고리즘에 대한 구현 및 성능평가 비교에 대해 논한다. 2장에서는 MSMP알고리즘과 RIFLE알고리즘에 대하여 설명하고 3장에서는 실험에 사용된 데이터에 대하여 설명한다.4장에서는 두 알고리즘의 실험 결과와 분석평가에 대하여 설명하고 5장에서 결론에 대하여 설명한다.

2. 서열 검색 알고리즘 구현

2.1 MSMP 알고리즘

MSNP알고리즘은 염기서열 데이터베이스 내의 염기 서열들로 얻은 Restriction Site Map과 query isolate에서 얻은 Restriction Pattern과의 유사성을 측정하여 가장 유사한 restriction site Map을 추출한다. query isolate와 염기서열 데이터베이스 내의 서열의 유사성의 척도는 restriction site map과 restriction pattern의 공통적인 restriction site들의 최대 점의 개수를 사용한다. 그러나 query isolate에서 얻은 restriction pattern은 순서에 무관하다. 즉, Enzyme을 query isolate에 적용하여 (12,7,4)와 같은 조각의 길이를 구했다면 이는 {(12,4,7),(4,7,12)(4,12,7)(7,4,12)(7,12,4)}와 같은 restriction map을 만들 수 있다. 이때 조각의 수를 n개라 하면 n!개의 다른 restriction map을 만들 수 있으며 이 문제는 NP-complete 문제로 이미 증명되었다. MSMP알고리즘은 이 문제를 해결하기 위해 한 변수를 제한하여 문제를 해결하며 MSMP는 다음과 같은 과정으로 계산된다.

1) query isolate q의 restriction pattern과 염기서열 데이터베이스 D 내의 각 서열 d들에 대한 restriction site map을 얻는다.
2) q의 restriction pattern과 d의 restriction site map들을 비교하여 유사도를 측정한다.
3) 이 유사도를 이용 query isolate q에 대한 생물학적 정보를 추론한다.

이때 유사도의 측정은 다음과 같이 정의할 수 있다.

1) restriction site map M(d)와 restriction pattern P(q)로부터 얻어진 n!개의 절단 위치 지도를 비교하여 공통된 절단 위치들의 개수 common(M(d),P(q))를 계산한다.
2) 최대 공통 절단 위치의 개수 maxcommon(M(d),P(q)로 정의한다.
3) restriction site map내의 site의 개수와 restriction pattern내의 site 개수중 큰 수를 maxsite(M(d),P(q))로 정의한다.
4) 유사도의 측정은 k=maxsite(M(d),P(q)) - maxcommon(M(d),P(q))로 정의한다.

MSMP는 query isolate의 절단 위치의 수를 n이라 하고 염기서열 데이터베이스 내의 서열들의 restriction site의 수를 m이라고 할 때 $|n-m| \leq 2$로 제한하여 NP-complete 문제를 해결한다.

2.2 RIFLE 알고리즘

RIFLE알고리즘은 두 개의 restriction pattern에 대한 optimal distance value를 산출하여 유사성 정도를 측정한다. Restriction pattern은 Restriction map과 restriction profile로 구분된다. restriction map은 이미 서열을 알고 있는 염기서열 데이터베이스에서 얻은 restriction pattern이기 때문에 Enzyme에 의해 절단이 되었더라도 단편들의 순서는 정해져 있다. 반면 restriction profile은 서열을 알지 못하는 새로운 유기체에 대해서 실험으로 얻어진 길이정보로서 단편들의 순서는 알 수 없다. REFLE은 restriction maps과 restriction profiles에 대해 유사성을 측정한다. 즉 REFLE은 순서가 있는 restriction map을 순서 없이 알고리즘에서 정렬 사용함으로서 restriction profile로 사용한다.

RIFLE은 Dynamic programming[5]을 기반으로 한다. 이 Dynamic programming은 최적인 유사서열을 찾는데 사용되며 sequence comparison과 sequence alignment에 많이 쓰인다. Dynamic programming을 사용할 경우 문제 해결을 위해 score matrix가 필요하지만, RIFLE은 단편의 길이에 대한 계산임으로 score matrix가 필요 없다. RIFLE은 다음과 같은 과정에 의해서 계산된다.

1) restriction profile s와 restriction map t를 입력받는다.
2) 두 서열 s와 t에 대해서 Dynamic programming을 적용한다.
3) score 함수 fi 은 $f(x,y) = |x-y|$이고 $f(x,0) = x$ 이다.
4) 다음과 같이 restriction profile과 restriction map이 주어졌을 때
restriction profile s=(23,17,5)
restriction map t=(24,17,10,5)
두 서열의 거리 $Df(s,t) = 11$이다.

3. 실험 데이터의 생성 방법

실험 데이터는 restriction pattern 1200개와 restriction site map 1200를 대상으로 하였다.

3.1 데이터베이스

실험 데이터베이스는 http://rdp.cmd.msu.edu/downlo-ad/SSU_rRNA/alignments에서 얻은 Ribosomal Data Project(RDP)를 사용하였다. RDP는 16s rRNA 유전자들로 구성된 염기서열 데이터베이스이다. 이 데이터베이스 내의 서열들은 계통학적 발생관계(Phylogenetic relatedness)가 알려져 있다. 실험에서는 서열들 중에서 Bacteriol부분의 2000개 정도를 실험에 사용하였다. 이 서열들은 공통 조상들로부터 파생되었기 때문에 서열간의 유사도가 높다. 서열들은 GenBank형식으로 되어 있으며 본 논문에서는 실험을 목적으로 서열들에서 ORF(Open Reading Frame)부분을 추출하여 FASTA형식으로 전환하였다.

3.2 DB서열의 길이정보

제한효소를 사용하여 서열들을 절단한 다음 각 조각 패턴들의 길이 정보만을 수치 값으로 변환하여 새로운 길이 정보 데이터베이스를 구축하였다. 실험에 사용된 제한효소는 Hpa II, Hinf I, Rsa I, Hha I이며 이 제한 효소들은 다음의 특성을 가진다.

Hpa II - C / CGG
Hinf I - G / ANTC
Rsa I - GT / AT
Hha I - GCG / C

각 제한효소들은 서열에서 일치하는 부위를 찾아 '/' 부분을 절단한다. 제한효소는 하나 또는 여러 개를 서열에 적용하였으며 각각의 위치는 Sliding Window 방식을 사용하여 구현하였다.

3.3 query 서열 길이정보

실험에 사용된 query서열 길이 정보는 데이터베이스 내의 서열들로부터 얻어진 restriction site map에서 각각의 단편들에 0~5%의 랜덤 오차를 적용하여 생성하였다.

4. 실험 결과 및 평가

표 1은 MSMP를 이용하여 각 query가 검색한 데이터베이스 서열들로 실험데이터 1200개중 자신을 제외한 1199개의 서열과 검색한 결과의 일부분이다.

env.OPB13 : 2
(AF018192, 12, 98%) (AF018196, 12, 98%)
AF018192 : 2
(env.OPB13, 12, 98%) (AF018196, 12, 99%)
AF018196 : 2
(env.OPB13, 12, 98%) (AF018192, 12, 99%)
Aqu.pyroph : 0

AF068791 : 1
(AF068801, 15, 99%)
AF068801 : 0
AF068807 : 0
Hbg.subter : 0

표 1 MSMP 결과

제일 상단의 env.OPB13 : 2는 env.OPB13이라는 이름을 가진 query로 데이터베이스를 검색한 결과 두 개의 서열이 검색되었다는 의미이다. 검색된 데이터는 서열이름,단편 개수, 실제 유사도 형태로 출력하였다.

표 2는 RIFLE을 알고리즘을 이용해서 Distance로 Rank를 적용한후 Rank 20까지의 검색한 결과이며 이 서열들이 실제 유사도를 비교하여 얼마나 신뢰성 있는 검색을 하는지를 나타낸다.

Rank	Hpa II Name	Dist	Frag	Sim	Hpa II, Hinf I, Hha I, Rsa I Name	Dist	Frag	Sim	Similarity Name	Sim
env.OPB13										
1	env.OPB13	14	13	100	env.OPB13	19	17	100	env.OPB13	100
2	AF018192	14	12	98	AF018192	19	17	98	AF018192	98
3	AF018196	14	12	98	AF018196	147	16	98	AF018196	98
4	AJ240990	159	12	71	env.WCHB25	150	24	68	AF068801	88
5	Mlb.organ3	164	11	71	AB15887	163	18	66	AF068807	86
6	AB015560	177	11	71	Hp.foetida	166	17	68	AF68791	87
7	Y14818	184	11	73	Mlb.organ2	167	17	71	Aug.pyrophl	78
8	Sch.teran2	184	13	70	AF068807	168	18	69	Hdg.subterL	76
9	Srh.terang	184	13	70	AF033558	168	18	70	str.BM_17	75
10	Z78203	184	13	70	AF056343	169	18	66	env.OPB_5	74
11	env.RB_40	191	10	69	Mlb.radtol	171	18	71	env_OPB33	74
12	AJ231189	193	10	67	Mlb.spE75	171	18	71	AF068788	74
13	U58018	194	10	71	Mlb.mesophl	171	18	72	env.OPB45	73
14	Mlb.spF48	194	10	72	Mlb.spE18	171	18	72	AJ009501	73
15	Mlb.extor4	194	10	72	D.radiodur	172	17	60	Tdv.TGE_P11	73
16	Mlb.GK101	194	10	72	Blb.Bf15	173	18	71	AJ237665	73
17	Mlb.rhodi2	194	10	73	Blb.PC3039	177	18	70	Tt.subterrnl	73
18	Mlb.GK118	194	10	71	AJ009501	180	16	73	Cum.prtlytl	73
19	AJ009481	198	11	67	AJ009451	181	18	73	Rht.marin21	73
20	AJ009451	200	12	73	U81656	182	17	67	Rht.marinl	73

표 2 RIFLE 결과

실험결과 RIFLE과 MSMP 모두 자신의 서열을 정확하게 찾아준다. 그 다음 문제는 자신의 서열을 제외한 나머지에서 실제 유사도가 높은 서열을 얼마나 잘 찾을 수 있는가에 대한 문제로 이를 위해 RIFLE과 MSMP알고리즘의 실험 결과를 통합하였다.

#	similarity name	%	RIFLE name	dist	frag	%	MSMP q.dl(0.0) name	frag	%	q.dl(1.0) name	frag	%	q.dl(0.1) name	frag	%
AF068791(15)															
1	AF068791	100	AF068791	22	15	100	AF068791	15	100	AF068791	15	100	AF068791	15	100
2	AF068801	99	AF068801	24	15	99	AF068801	15	99	AF068801	15	99	AF068801	15	99
3	AF068807	98	T.spTok8	198	14	71							AF068807	14	98
4	AF018196	88	D.geothe2	190	16	70									
5	AF018192	89	D.geothe3	190	16	70									
6	env.OPB13	87	D.geothe4	190	16	70									
7	Aqu.pyrophl	73	D.geothe5	190	16	70									
8	Hdg.subterH	73	str.BM_17	192	14	76									
9	str.BM_17	79	AF020705	194	15	73									
10	Tt.maritiml	78	T.spNMX2	194	15	74									
11	Tt.subterrnK	75	AF068807	204	14	98									
12	env.OPB45	75	Tt.subterra	208	16	75									
13	env.OPB14	74	T.spTok20	212	15	73									
14	T.thmoph27	74	T.spTok3	212	15	72									
15	T.thmoph137	74	T.spW38	212	15	73									
16	T.spFjs3AJ	74	T.spYSPID	220	17	73									
17	T.spNMX2	74	T.aquaticu	220	17	73									
18	T.spH2	74	T.spCHGI	223	16	73									
19	AB020889	74	env.OPB80	224	17	71									
20	T.flavus2	74	env.OPB80	224	17	71									

표 3 MSMP, RIFLE 결과 비교 분석표 I

#	similarity name	%	RIFLE name	dist	frag	%	MSMP q.dl(0:0) name	frag	%	q.dl(1:0) name	frag	%	q.dl(0:1) name	frag	%
AF018192															
1	AF018192	100	AF018192	22	17	100	AF018192	17	100	AF018192	17	100	AF018192	17	100
2	AF018190	99	env.OPB13	23	17	99	env.OPB13	17	98	env.OPB13	17	98	env.OPB13	17	99
3	env.OPB13	98	AF018190	150	18	99							AF018190	16	99
4	AF068801	89	env.WCHB25	155	24	68									
5	AF068791	83	AJ009501	171	18	73									
6	AF068807	88	AF033558	171	18	71									
7	Aqu.pyrophl	78	AF068807	171	18	69									
8	Hdg.subterL	78	D.radiodur	175	17	69									
9	str.EM_17	75	AF068801	197	19	69									
10	AJ237665	73	Y13595	199	19	71									
11	Tt.subterrnl	73	T.sp_ac2	199	19	71									
12	Tt.maritiml	73	AF068791	215	13	89									
13	Cm.prtlytl	73	D.murrayi2	215	17	69									
14	env.OPB45	73	D.radiodu2	217	18	69									
15	AJ009501	73	D.prtlytic	219	19	69									
16	Hsp.moscovl	73	D.murrayi3	225	18	60									
17	AJ224039	73	D.murrayi1	225	18	63									
18	AJ224041	73	AJ237665	225	18	73									
19	Tdv.TGE_P11	73	AJ224041	227	19	72									
20	AF018195	72	str.EM_17	257	19	75									

표 4 MSMP, RIFLE 결과 비교 분석표 II

비교 결과 표3에서 보면 AF068791과 유사한 서열이 3개가 있다. 하나는 자신이고 다른 두 개는 98%로 상당히 유사하다. 반면 유사도가 높은 AF068807서열(유사도 98%)의 경우 RIFLE에서는 RANK가 11로 밀려난 것을 볼 수 있다. 이 경우 실제 유사도를 측정해 보지 않은 상태라면 유사도가 현저하게 떨어지는 다른 서열이 마치 AF068807서열보다 더 유사하다고 착각할 수 있다. MSMP의 경우도 RIFLE과 마찬가지로 AF068807서열을 찾지 못했다. 하지만 MSMP의 경의 Merge를 허용한다. restriction site map에 한번의 Merge를 허용하였을 경우 AF068807서열을 유사한 서열로 찾을 수 있었다.표2에서는 RIFLE에 의한 결과에서 RANK 4를 기록한 env.WCHB25 서열은 AF018192서열과는 restriction pattern의 개수가 많이 차이난다. 실제 검색 대상인 AF018192서열은 17개의 단편을 가지고 있고 env.WCHB25서열은 24개의 단편을 가지고 있다. MSMP에서는 NP-complete 문제를 해결하기 위해서 단편의 차수가 2 이하인 서열을 대상으로 검색하는데 반면 RIFLE은 이러한 고려사항을 가지고 있지 않다.100개의 서열데이터 대하여 여러 가지의 제한효소를 적용하였을 경우 MSMP의 경우 결과는 표5와 같다. 반면 RIFLE의 경우 D.radiopug 서열만이 MSMP에 비해 유사도가 높은 서열로 RANK되었다.

Hpa II	HpaII|Hinf1	Hpa2|Hinf1 |Rsa1	Hpa2|Hinf1 |Rsa1|Hha
12/100	15/100	11/100	5/100

표 5 MSMP의 서열 검색 결과

5. 결과

본 논문에서는 서열을 알지 못하는 한 유기체에 대하여 제한 효소를 이용하여 단편의 길이 정보를 구하여 단편 길이만으로 유사성 검색을 하는 두 알고리즘에 대해서 분석을 하였다. query 서열과 데이터베이스 내의 서열에 대하여 실험에 편리한 형태로 가공을 하였으며 가공된 데이터로 MSMP와 REFLE 알고리즘을 구현해 각각 적용하여 결과를 산출하였다. 실험결과 RIFLE은 Dynamic Programming을 기반으로 한 알고리즘이기 때문에 시간 복잡도는 $O(n^2)$이다. 이에 반하여 MSMP는 NP-complete문제로 보고 이를 해결하기 위해 단편의 차수를 제한하는 휴리스틱을 적용하였다. 수행속도 면에서는 RIFLE이 빠르지만 RIFLE이 산출해낸 결과는 MSMP가 산출한 결과보다 신뢰성이 떨어진다. 그리고 MSMP는 사용자의 결정에 따라 Error Bound와 Merge의 적절한 수치를 입력함으로써 실제 유사도가 높은 신뢰성 있는 서열만을 검색해 낼 수 있었다.

참고문헌

[1] http://www.ncbi.nlm.nih.gov

[2] Pearson, W. R. "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms." Genomics, 11 (3), 635-50. 1991

[3] Hermjakob, H and Giegerich, R. and Arnold, W "RIFLE: Rapid Identification of Microorganisms by Fragment Length Evaluation" AAAI Press Menlo Park, CA, USA P:131-137

[4] Jin Kim, James R. Cole, and Sakti Pramanik "Inferringrelatedness of a Macromolecule to a sequencedatabase without sequencing" Nucl.Acids Res, Vol.12 pp.175-180, 1984

[5]Pearson, W. R. and Miller, W. Dynamic programming algorithms for biological sequence comparison. Methods Enzymol, 210, 575-601. 1992