[SVII-4]

BLAST/FASTA 를 활용한 미생물 유전체 비교용 도구의 개발

태홍석, 박기정*

(주)스몰소프트 정보기술연구소

미생물 유전체 프로젝트의 결과인 유전체 서열에 대해, 비교 유전체 분석을 수행할 수 있는 분석 도구인 GComp 를 개발하였다. 이 도구는 국부 상동성 계산을 BLAST 나 FASTA 를 사용하여 수행한 후에 이 결과를 파싱하여, 상동성을 보이는 부분을 분석하고 mapping 하여, 두 유전체간의 상동성 정도를 일목요연하게 보여줄 수 있는 테이블과 파일들을 생성한다. 한편, 그 결과를 그래픽으로 표시하고 전체를 살펴볼 수 있는 인터페이스 기능을 구현하였다. 시험 데이터로 기존의 미생물 유전체 서열을 상대로 분석하면서, 유전체 서열의 핵산 및 단백질 수준에서의 비교결과를 통해 두 유전체에 대한 비교 정보를 효과적으로 확인할 수 있었고, 보다 다양한 분석을 위한 추후의 개발 방향을 설정할 수 있었다.

서 론

유전체 프로젝트가 활발히 진행되면서, 유전체 수준에서 유전자의 기능 정보를 분석하고 비교하는 방법이 일반화되어 가고 있다. 한 유전체에 대한 분석 결과는 그 유전체에 포함된 유전자의 위치와 예측되는 기능들의 집합으로 표시된다. 이러한 유전체에 대한 정보는 유전자 집합의 기능에 대한 위치별 리스트나 gene ontology 에 관한 데이터베이스의 클러스터별로 정리하여 표시하거나, 가시화 도구를 사용하여 원형의 유전체 지도 혹은, 선형의 유전체 지도로 표시한다. 현재는 이러한 유전체 분석 결과는 해당 홈페이지를 통해 인터넷에서 제한된 정보를 살펴볼 수 있는 정도이고, 이러한 정보를 생성하거나, 보다 효율적으로 이들 정보를 사용하기 위해서는 고가의 상용도구를 사용해야만 하고, 이러한 상용 프로그램들도 현재는 개발 단계에 있다.

유전체 프로젝트의 중요한 목적이자 방법 중의 하나는, 대상이 되는 생명체의 유전체를 기존의 다른 유전체와 비교하여, 관련 유전자의 종류와 유사정도 등을 통해 유전자의 기능에 대한 포괄적인 정보를 구하는 것으로, 일반적으로 비교 유전체학(comparative genomics)하고 한다. 하나의 유전체에 대한 자체 분석이 완료되거나 진행 중일 때, 일반적으로 생물학자들이 가장 원하는 분석요구의 하나가 바로 유사하거나 관심 있는 다른 생명체와의 유전체 정보 비교분석이다. 유전체 정보를 비교하는 것은, 유전체에 대한 정보를 표시하는 방법의 표준화를 필요로 하는 일이나, 현재 유전체 정보의 내용과 표시 방법이 임의적이므로, 비교 유전체 분석의 내용과 방법은 대단히 유동적일 수밖에 없다.

유전체를 유전체 핵산서열로 비교하는 방법과 유전체 상의 단백질 서열로 비교하는 것을 가장 일반적으로 생각해 볼 수 있는 방법인데, 실제로 이를 두 유전체의 비교를 위해 사용하고 있

다(3,4). 현재 가장 많이 알려져 있는 도구로 TIGR 에서 개발한 MUMmer(5,6)가 있으며, 핵산 비교를 위한 NUCmer 와 단백질 비교를 위한 PROmer 로 구성된다. 단백질 비교는 예측된 단백질 서열에 대한 것이 아닌 임의로 6-frame 으로 변환한 서열에 대한 것으로 유전자 예측 프로그램이나 그 기능과는 무관하다. 이 프로그램들의 장점으로 주장되는 것은, 빠른 처리 속도로 이는 서열비교 알고리즘으로 cross match 와 같은 방식을 사용하며 속도 개선을 위해 suffix tree 를 사용했기 때문이다. 한편, 이러한 프로그램들에 대한 가시화 프로그램에 대한 연구들도 진행되고 있으며(7), MUMmer 프로그램들의 처리 결과를 가시화하여 나타내기 위해 별도의 프로그램인 DisplayMUMs 가 개발되었다. 이 프로그램들의 실행환경은 UNIX 이며, 현재 Windows 환경에서의 버전은 개발되지 않았다.

본 논문에서는, 미생물 유전체 프로젝트의 결과인 유전체 서열을 입력으로 하여, 두 유전체를 핵산이나 단백질 서열의 수준에서 비교하는 비교 유전체 계산과 그 인터페이스를 포함한 기능을 하나의 도구로 통합하여 수행하는 프로그램으로 GComp 를 개발하였다. 생물학 연구자의 일반적인 실행환경을 고려하여 VisualC++을 사용하여 윈도우즈상의 실행 프로그램으로 개발하였으며, 기존의 동일 목적 프로그램들의 기능을 반영하도록 설계하고 구현하였다. 상동성 비교의 효율을 위해 MUMmer 에서 사용하는 알고리즘 대신 BLAST(1,2)와 FASTA 프로그램을 활용하여, 그 결과를 비교에 반영하였다.

이 프로그램의 시험을 통해 여러 미생물 유전체에 대한 비교를 수행하여, 핵산과 단백질 수준의 비교에 대한 뚜렷한 차이를 볼 수 있었으며, 두 가지 비교에서 모두 일반적으로 예상할 수 있는 것보다 훨씬 낮은 수준의 유사성이 미생물의 유전체에서 보이는 것을 알 수 있었다. 기능에 대해 보다 직관적인 이해를 얻기 위해, 다양한 자료의 산출과 인터페이스의 개발이 추후 연구를 통해 이루어져야 할 것이며, 현재의 gene ontology 에 대한 이해가 확대되면서 이에 대한 클러스터 정보 등이 프로그램에 포함되어야 할 것이다.

시스템 구현과 방법

두 개의 유전체 서열을 국부상동성 수준으로 비교하기 위해 BLAST 나 FASTA 를 실행한다. BLAST 나 FASTA 에 의한 상동성 분석 결과를 파싱하여, 일정 수준 이상의 유사성을 보이는 것을 상동성으로 판정하여, 이들에 대한 위치와 표시 리스트를 작성한다. 작성된 리스트를 그래픽으로 표현한다.

국부 상동성 비교

두 유전체를 전체 유전체 서열 차원에서 비교하기 위해 BLAST 를 이용하였다.

두 개의 유전체 서열 중 완성된 유전체 서열을 가진 하나의 유전체 서열에 대해 blast 의 formatdb 를 이용해서 BLAST DB 로 구축하고, 다른 하나의 유전체 서열의 완성된 유전체 서열 또는 contig 들을 질의어로 해서 blastn 계산을 실시한다. 단백질 수준의 비교를 위해서 tblastx 를 사용한다. 두 가지의 경우 별도의 파싱 프로그램을 사용한다. 한편, 상동성 비교 알고리즘에 대한 분석 결과를 비교하기 위해 BLAST 와는 다른 알고리즘을 사용하는 프로그램으로 FASTA 를 사용하였다. 이 경우에도, 두 개의 유전체 서열 중 완성된 유전체 서열을 가진 하나의 유전체 서열

에 대해 FASTA DB 로 구축하고, 다른 유전체의 완성된 유전체 서열 또는 contig 들을 질의어로 해서 FASTA 계산을 실시한다. FASTA 의 경우에는 현재, 질의어 서열의 크기가 20Kb 정도로 제한되어 있어 BLAST 와의 계산 결과 비교용으로만 사용하였다. FASTA 결과의 대한 파싱은 BLAST 와는 또 다른 파싱 프로그램을 사용하여 분석한다.

상동성 리스트 작성

blastn 의 결과를 처리하는 경우를 예로 상동성 리스트 작성 과정을 설명한다. blastn 의 output file 을 파싱하여, 특정 유사성 이상을 나타내는 (E-value threshold 값을 기준) 두 서열 사이의 모든 정렬들을 링크드 리스트에 연결한 후 각 정렬을 대상 유전체 서열에 mapping 한다. Contig 들이 질의어로 입력되면 각 contig 별로 정렬의 링크드 리스트가 따로 작성한다.

하나의 contig 가 유전체의 여러 군데 mapping 이 되는 경우를 처리하기 위해, 정렬을 그룹으로 묶어서 처리하였다. (그룹은, 하나의 contig 에 대해서, 동일하거나 특정 거리 내에서의 유사한 위치에 mapping 되면서 같은 방향을 갖는 정렬들로 구성된 것이다). 그룹 형성의 특정 거리 값은 option 으로 설정할 수도 있지만, 현재는 특정 거리를 각 contig 의 길이로 두고 처리하도록 하였다.

각 contig 의 blastn 파일을 파싱하여, 해당 contig 에서 E-value 가 가장 낮은 정렬을 선택하고 그 정렬의 position 을 기준으로 앞 방향과 뒤 방향 각각 contig 의 길이 범위 내에 위치하는 정렬들의 집합을 하나의 그룹으로 포함시키면서 링크드 리스트에 연결한다. Mapping 되는 start position 과 end position 을 각각 node 로 두어, position 에 따라 각 node 를 연결한다. 즉, 정렬을 그룹에 포함시킬 때 마다 정렬의 start position 과 end position 인 두 node 들은 이 링크드 리스트의 node 들을 순차적으로 검색해서 알맞은 위치로 끼어 들어간다. 한 contig 에서 default 로 3 개의 그룹을 형성할 수 있고, 첫 번째 그룹의 기준이 되는 정렬은 그 contig 에서 발생한 정렬중에서 가장 E-value 가 낮은 정렬이 항상 선택되는 반면, 두 번째와 세 번째 그룹의 기준이 되는 정렬은 그 이전 그룹들에 포함 되지 않으면서 정렬 E-value 가 threshold 값 이하 되는 정렬이 선택된다. 이 기준에 만족하는 정렬이 없을 경우에는 더 이상의 그룹은 형성되지 않는다. 한 contig 에서 형성할 수 있는 그룹은 option 으로 9 개까지로 설정할 수 있고, 그룹의 기준이 되는 정렬의 E-value 의 threshold 값은 option 으로 설정할 수 있다.

tblastx 와 FASTA 의 결과에 대한 파싱과 정렬 링크드 리스트도 유사한 방법으로 생성된다.

이러한 처리 결과로, 하나의 contig 에 대해서 position 으로 sorting 된 다수의 그룹의 정렬 링크드 리스트가 생성된다.

유전체 상동성 가시화 표현

각 그룹의 정렬 list 에 대해, 순차적으로 정렬에 대한 start position 과 end position 사이의 부분에 대해 DB 서열과 align 되는 도형(covered)으로 표시하고 그룹의 범위 내에서 어떠한 정렬이 나타나지 않는 부분은 align 되지 않는 부분(uncovered)으로 표시한다. covered 부분에서 overlap 되는 여러 정렬들에 대해서는 E-value 가 낮은 정렬을 우선적으로 display 한다.

각 contig 에 대해 정렬의 리스트를 별도의 윈도우를 통해 출력할 수 있는 기능을 구현하였다. 한편, 전체 contig 들에 대한 리스트를 하나의 텍스트 파일을 통해 출력할 수 있는 기능도 구현하였다.

결과 및 고찰

개발 프로그램의 기능과 사용 환경

Linux 상에서 실행된 BLAST 나 FASTA 계산 파일의 결과와 대상이 되는 유전체 서열의 FASTA 파일을 입력으로 하여, 정렬 list 생성과 그래픽 표현을 위한 계산 프로그램을 Windows 상에서 실행한다.

입력할 파일들의 위치를 정해진 지시에 의해 선택하면, 이들을 모두 읽어 들여 Fig. 1 과 같이 대상 유전체에 대한 covered/uncovered 그래픽을 출력한다.

윈도우의 윗 부분에 그래픽 출력에 대한 보조 데이터가 표시되는데, scale 을 표시하는 값이나, 전체 유전체 길이 및 covered 부분의 길이의 총합과 전체에 대한 covered 의 길이 비율 등이 표시되고, scale 값을 변화시켜 그래픽 화면을 재출력할 수 있다.

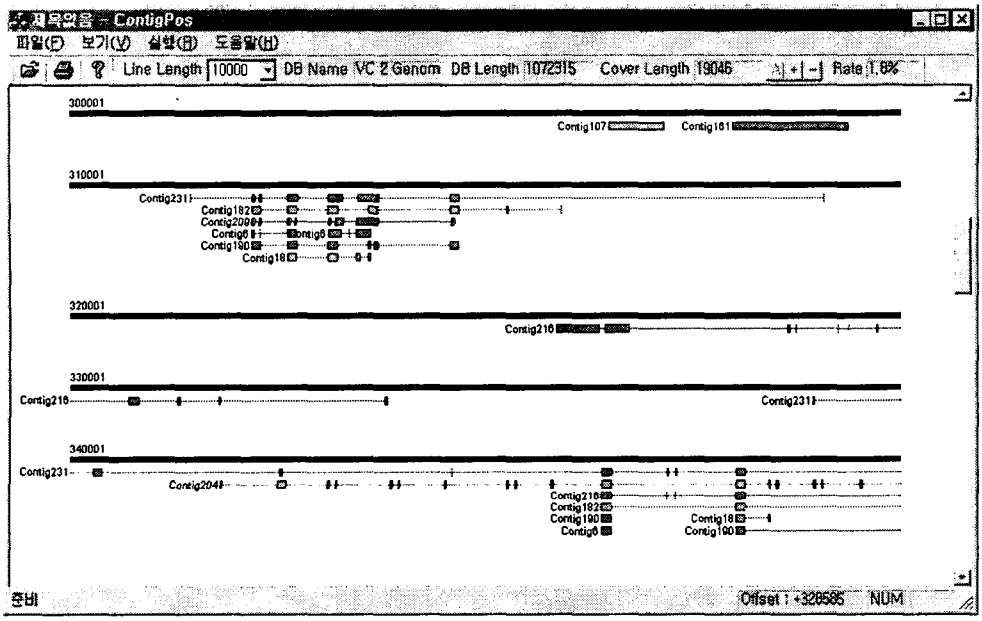
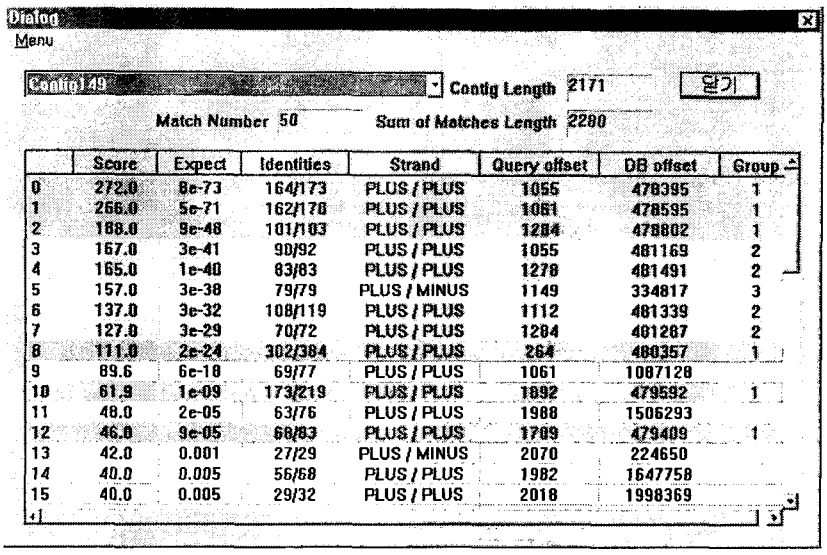


Fig. 1. 비교 유전체 mapping 그래픽 표현 화면.
파란색 사각형 : forward 방향의 covered
빨간색 사각형 : reverse 방향의 covered
파란색/빨간색 직선 : 그룹 연결선으로 uncovered



	Score	Expect	Identities	Strand	Query offset	DB offset	Group
0	272.0	8e-73	164/173	PLUS / PLUS	1055	478395	1
1	266.0	5e-71	162/178	PLUS / PLUS	1081	478595	1
2	188.0	9e-48	101/103	PLUS / PLUS	1284	478802	1
3	167.0	3e-41	90/92	PLUS / PLUS	1055	481169	2
4	165.0	1e-40	83/83	PLUS / PLUS	1278	481491	2
5	157.0	3e-38	79/79	PLUS / MINUS	1149	334817	3
6	137.0	3e-32	108/119	PLUS / PLUS	1112	481339	2
7	127.0	3e-29	70/72	PLUS / PLUS	1284	481287	2
8	111.0	2e-24	302/384	PLUS / PLUS	264	480357	1
9	89.6	6e-18	69/77	PLUS / PLUS	1061	1087128	
10	61.9	1e-09	173/219	PLUS / PLUS	1892	479592	1
11	48.0	2e-05	63/76	PLUS / PLUS	1988	1506293	
12	46.8	9e-05	66/83	PLUS / PLUS	1709	479409	1
13	42.0	0.001	27/29	PLUS / MINUS	2070	224650	
14	40.0	0.005	56/68	PLUS / PLUS	1982	1647758	
15	40.0	0.005	29/32	PLUS / PLUS	2018	1998369	

Fig. 2. Contig 별 정렬 리스트 출력.
동일한 색깔은 같은 그룹을 표시한다.

한편, 각 contig 에 대한 정렬 list 의 자세한 내용을 보기 위해 Fig. 2 와 같은 contig 별 정렬 list 테이블을 출력할 수 있다. 이 출력 화면의 최상단에서 contig 를 지정할 수 있으며, 화면 상단에 각 contig 에 대한 정보로, 길이, 정렬의 개수, 정렬 길이의 합이 출력된다. 리스트에서는 각 정렬에 대한 정보로, 정렬 score, identity, 질의어 상에서의 위치와 대상 유전체 서열 상에서의 위치 및 그룹 번호 등이 출력된다.
전체 정렬에 대한 정보를 하나의 파일로 출력하기 위해, Fig. 3 과 같은 텍스트 파일을 생성할 수 있다.

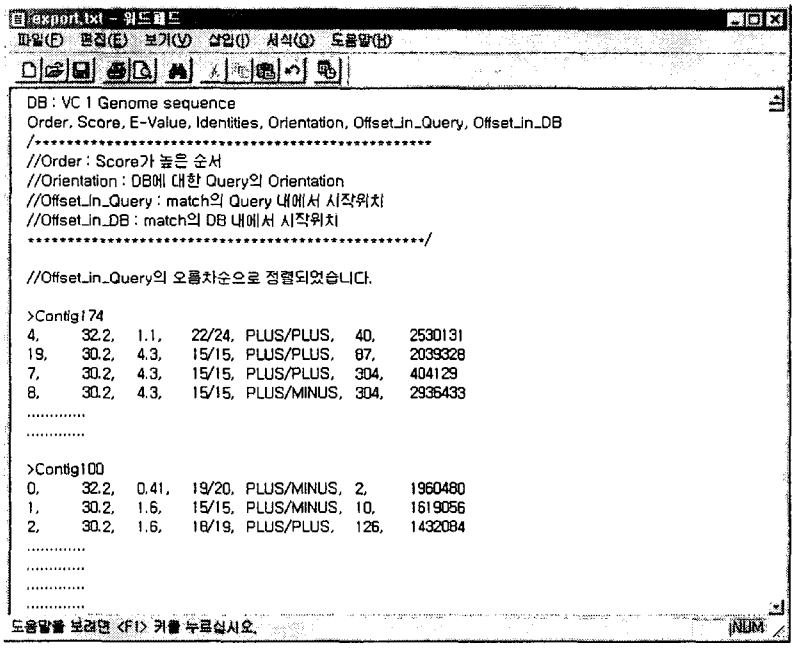


Fig. 3. 전체 contig 들에 대한 전체 정렬 리스트 출력 텍스트 파일.

유전체 데이터 분석 사례 연구

개발된 프로그램의 시험과 실제의 미생물 유전체에 대한 비교 유전체 데이터 생성을 위해, *Bacillus halodurans* C-125(이하 BH), *Bacillus subtilis*(이하 BS), *Escherichia coli K-12 MG1655*(이하 K12), *Escherichia coli* O157:H7(이하 O157), *Vibrio cholerae*(이하 VC), 그리고 현재 아직 contig 단계에서 진행 중인 미생물의 유전체(이하 CS)를 대상으로 계산을 수행하였다. 비교 쌍은 BH_BS, K12_O157, CS_VC, BS_K12 로 하였고 cut off E-value 는 10 을 default 로 하였다.
BH_BS 비교에서 BS sequence length 인 4214814 bp 중 45.4%인 1913524 bp 가 BH 에 의해서 cover 되는 것으로 나타났다. 높은 E-value 를 가지는 정렬들이 cover rate 에 영향을 주는 정도를 알아보기 위해서 cuff off E-value 를 다양하게 설정해 보았다. cut off E-value 를 10, 1.0, 0.1, 0.01, 0.001 로 했을 때 각각의 cover rate 는 45.4%, 45.2%, 44.7%, 44.3%, 44.0%로 나타났다.

Table 1. BH_BS 비교 쌍에 대한 GComp 계산 결과 (blastn 사용).

Cut off E-value	10	1	0.1	0.01	0.001
유효한 정렬 개수	6864	6335	5419	4901	4567
highest E-value	4.1	0.45	0.05	0.006	6e-04
total cover length (bp)	1913524	1904254	1883114	1867176	1856073
cover rate (%)	45.4	45.2	44.7	44.3	44.0

이와 같은 조건으로 시험에 사용된 비교 쌍에 대해 같은 계산을 수행하였다 (Table 2).
K12_O157 비교 쌍은 79.5%의 cover rate 를 보여준다. forward 정렬 그룹만의 cover rate 가 74.0% reverse 정렬 그룹만의 cover rate 가 14.9%를 나타낸다.

CS_VC 비교 쌍은 CS 의 contig file 들을 질의어로 입력했다. VC 의 경우 Chromosome 이 둘이기 때문에 Chromosome 1 과 Chromosome 2 를 따로 비교해 보았다. Chromosome 1 에서는 Chromosome 길이 2961149 bp 에 대해 12.0%의 cover rate 를 나타내었고, Chromosome 2 에서는 Chromosome 길이 1072315 bp 에 대해 15.0%의 cover rate 를 나타내었다.

K12 유전체의 길이는 4639221 bp 이며, BS 유전체의 길이는 4214814 bp 이다. BS 를 DB 로 구축하고 K12 를 질의어로 한 K12_BS comparison 의 경우 cover rate 는 19.2%로 나타났고, K12 를 DB 로 구축한 BS_K12 의 경우는 17.1%로 나타났다.

각 경우에 나타난 유전체 상동성 결과에서, 핵산 수준의 유전체 서열에서는 일반적으로 진화적인 차이 정도로 예측되는 수준이상으로 큰 차이가 나는 것을 볼 수 있었다.

Table 2. 비교 쌍들에 대한 GComp 계산 결과 요약 (blastn 사용).

비교 쌍	DB length (bp)	질의어 length (bp)	cut off E-value	cover length (bp)	cover rate (%)
BH_BS	4214814	4202353	10	1913524	45.4
			0.01	1867176	44.3
K12_O157	5528970	4639221	10	4397839	79.5
			0.01	4390460	79.4
CS_VC1	2961149	447076	10	354338	12.0
			0.01	297042	10.0
CS_VC2	1072315	447076	10	161079	15.0
			0.01	105050	9.8
BS_K12	4202353	4639221	10	795428	17.1
			0.01	700617	15.1
K12_BS	4639221	4202353	10	810089	19.2
			0.01	714055	16.9

상동성 분석 알고리즘에 대한 비교 유전체 분석 결과

FASTA 를 사용했을 때의 비교 결과를 BLAST 실행 결과와 비교해 보기 위해, K12_BS 비교 쌍을 대상으로 계산을 수행하였다. BS 의 유전체 서열로 FASTA DB 로 구축하고 K12 의 유전체 서열을 20Kb 단위로 나누어 다수 개의 질의어로 입력한 FASTA 의 결과를 blastn 결과와 비교해 보았다. FASTA 의 경우에는 질의어 길이에 제한이 있기 때문에 이러한 처리를 하였다. Display 하기 위해서 그룹을 형성할 때 cut off value 를 0.01 으로 두었다. blastn 의 결과 19.2%로 나타났지만 FASTA 의 결과는 12.7%로 나타났다. 이는 두 프로그램이 갖고 있는 알고리즘의 차이이기도 하지만, 파라메터를 동일 조건으로 둘 수 없는 차이이기도 하므로, FASTA 의 실행 결과에 영향을 주는 파라메터를 다양하게 바꾸면 BLAST 와 유사한 결과를 보이는 결과가 얻어질 수 있을 것으로 예측된다. 따라서, 비교 유전체 결과 분석이 파라메터에 대해 대단히 민감하게 다른 결과를 낼 수 있음을 예측할 수 있다. BLAST 의 경우, 파라메터에 대한 결과의 차이에 대한 분석이 추후 연구에서 이루어질 것이며, 이를 기반으로 비교 유전체 분석 방법의 다양화나 표준화 및 결과 해석에 대한 기준 등에 대한 정보를 구할 수 있을 것이다.

한편, 핵산과 단백질 서열의 상동성의 결과 차이점을 고려해서 BS 와 K12 의 서열을 대상으로 tblastx 로 상동성 계산을 처리하여 비교한 결과 blastn 보다 다소 높은 cover rate 를 나타내었다. GenBank 의 K12 유전체 데이터에서 coding region 으로 annotation 된 서열들은 K12 의 유전체의 87.9%를 cover 한다. K12 의 gene 들과 K12 유전체의 결과를 보면 많은 부분에서 여러 gene 들이 겹쳐서 나타난다. 세 곳 이상에서 높은 homology 를 보이는 gene 들도 다수 나타나며, IS1 protein InsB 의 경우 gene size 504 bp 인데, 6 곳에서 E-value 가 0.0 으로 나타났고, IS5 transposase 의 경우 1017 bp 이며 무려 11 곳에서 E-value 가 0.0 으로 나타났다.

한편, 위의 단백질 코딩 서열들을 amino acid 로 translate 하여 BS 를 대상으로 tblastn 을 이용해서 상동성 계산을 처리하여 비교를 수행한 결과, 47.1%라는 높은 cover rate 를 보이고 있다.

tblastx 나 tblastn 으로 전처리한 결과의 계산을 통해, 핵산 수준의 비교와 단백질 수준의 비교가 큰 차이를 보이고 있음을 볼 수 있는데, coding region 과 non-coding region 의 차이뿐만 아니라, coding region 에 대해서도 핵산 수준의 비교와 단백질 수준에서의 비교가 큰 차이를 보이고 있음을 나타내고 있다.

추후 연구

본 연구를 진행하고 그 결과를 통해, 초기에 설계했던 프로그램의 기능 외에 인터페이스와 분석의 종류 면에서 추후 연구를 통해 개발할 부분의 필요성과 방향에 대해 설계를 할 수 있었다. 특히, 비교 유전체를 위한 도구의 요구사항이 현재 뚜렷하지 않고 개발의 여지가 많아, 유전체 연구에 대한 심도 있는 분석을 통해 이러한 도구의 사양과 이를 해결하는 알고리즘에 대한 개발할 필요가 많을 것으로 생각된다.

BLAST 의 파라메터에 따른 분석결과의 차이점에 대한 추후 분석은, 비교 쌍의 특성에 대해 파라메터를 설정하는 방법과 일반적인 비교에서의 파라메터 표준화 내지 최적화에 대한 정보를 제공해 줄 것이다. Coding region 과 non-coding region 에 대한 별도의 분석과 구역별 상동성의 특징 분석을 통해, 유전체 비교 방법에 대한 일반적인 프로토콜과 표준 데이터에 대한 정보를 구할 수 있을 것이다. 한편, 이를 위해 gene prediction 프로그램과의 연계 및 통합을 통해 보다 효과적이고 통합적인 도구를 개발할 수 있을 것이다.

BLAST 나 FASTA 를 사용하기 위해 1 차적으로 Linux 서버에서 계산을 수행한 후에 Windows 용 GComp 를 사용해야 하므로, 대부분의 생물학 연구자에게는 사용 환경상 큰 제약이 생기게 되므로, Windows 인터페이스를 통해 Linux 상의 BLAST 와 FASTA 를 실행할 수 있는 기능을 개발하는 것이 주요한 사항일 것이다. 한편, 대상 유전체의 길이에 제한이 있는 FASTA 의 제한점을 간접적으로 보완하여 실질적으로 FASTA 를 비교유전체 분석에 활용하기 위해, 유전체 서열을 다수의 서열로 분할하고 이를 통합하여, 비교 유전체분석 결과에 반영하는 기능을 구현할 필요가 있을 것이다.

본 프로그램은 유전체 비교 분석 뿐 아니라 하나의 유전체에 대한 분석을 위해서도 활용될 수 있다. 즉, repeat region 이나, gene 의 분포나 gene 의 cluster 분포 등을 위해서 활용될 수 있을 것이므로, 이를 직접적으로 처리할 수 있는 기능을 구현하면, 보다 효과적이고 활용도가 높은 도구로 사용될 수 있을 것이다. 한편, 미생물 유전체 프로젝트 수행용 도구와 통합하여, 미생물 프로젝트의 수행 도중에 이 도구를 보다 효율적으로 활용할 수 있도록 할 수 있을 것이다.

참고문헌

1. **Altschul, S.F., W. Gish, W. Miller, E.W. Myers and D.J. Lipman**, 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

2. **Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman,** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids*

Research. **25**, 3389-3402

3. **Bansal, A.K., P. Bork, and P.J. Stuckey,** 1998. Automated pairwise comparisons of microbial genomes. *Math. Modelling and Sci. Computing*. **9**, 1-23.
4. **Bansal, A.K.,** 1999. An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics*. **15**, 900-908.
5. **Delcher, A.L., S Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg** 1999. Alignment of whole genomes. *Nucleic Acids Research*. **27**, 2369-2376.
6. **Delcher, A.L., A. Phillippy, J. Carlton and S.L. Salzberg,** 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*. **30**, 2478-2483.
7. **Folrea, L., C. Riemer, S. Schwartz, Z. Zhang, N. Stojanovic, W. Miller and M. McClelland,** 2000. Web-based visualization tools for bacterial genome alignments. *Nucleic Acids Research*. **20**, 3486-3496.

A genomics Tool for Microbial Genome Comparison Using BLAST / FASTA

Tae, Hongseok and Kiejung Park*

Information and Technology Institute, SmallSoft Co., Ltd., Daejeon 305-811, Korea

Abstract We have developed GComp as an analysis tool for comparative analysis of microbial genomes. The tool uses BLAST or FASTA as a preprocessing program for local alignments, parses the homology search results, and generates tables and files to show homology relationship between two genomes at a glance. The interface for graphical representation of the comparative genomic analysis has been also implemented. Through analysis of a few pairs of microbial genome sequences, the program has been proved to be practically useful and a few additional features have been devised and designed, which will be added in the further development.