김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제5호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 5 JUCHE106(2017).

세균 및 비루스게놈정보열람체계개발에 대한 연구

황진혁, 리석준

1995년에 처음으로 세균인 *Haemophilus influenzae*의 게놈배렬이 완전히 해석된 때로부터 현재까지 수많은 세균과 비루스의 게놈배렬이 완전히 해석되였으며 그 정보들이 인터네트에 공개되고있다.[1, 3]

우리는 본문방식의 세균 및 비루스게놈자료로부터 자료기지를 구축하고 망에서 리용할수 있는 게놈정보열람체계를 개발하기 위한 연구를 하였다.

재료 및 방법

NCBI(ftp://ftp.ncbi.nlm.nih.gov)에서 본문형태로 제공해주는 세균게놈정보화일과 Genbank 197.0의 gbvrl에 포함된 비루스게놈정보화일을 리용하였다.[2] 총 2 584종의 4 852개에 달하는 세균게놈자료와 102 229종의 1 433 313개에 달하는 비루스게놈자료를 리용하였다.

자료기지관리체계로서 MySQL 5.0을, 프로그람작성언어로서 Java(JDK 1.7)를, 봉사기용언어로서 PHP 4.2를 리용하였으며 홈페지우에서의 도형그리기에는 HTML 5.0에서 제공하는 기능을 리용하였다.

결과 및 론의

1) 세균 및 비루스게놈자료기지구축

기초자료로 리용한 세균 및 비루스게놈자료는 본문방식의 자료로서 일정한 형식으로 게놈정보들을 서술하고있다.

1개 세균종에 해당한 게놈정보는 여러개의 화일들에 나누어져있다. 표 1에 *Acetobacter pasteurianus* 게놈정보를 구성하는 6개의 화일과 그 내용을 주었다. 세균게놈정

표 1. *Acetobacter pasteurianus*게놈정보화일목록

No.	화일이름	화일의 내용	화일크기
1	AP011163.rpt	자료기지에서 해당 게놈의 등록정보	1KB
2	AP011163.gbk	해당 게놈의 구성정보, 참고문헌정보, 유전자정보 등을 포함하고있는 기본화일	3.5MB
3	AP011163.gff	게놈령역별유전자배치도	500KB
4	AP011163.ptt	게놈내에 존재하는 단백질정보	400KB
5	AP011163.rnt	게놈내에 존재하는 RNA정보	50KB
6	AP011163.fna	게놈전체 염기배렬	1MB

보를 구성하는 화일의 이름들은 모두 NCBI에 등록된 게놈배렬의 등록번호(AP011163)와 일치하며 각이한 확장자를 가진 6개의 화일로 구성되여있다. 다른 세균들의 게놈정보도 이와 류사한 방식으로 구성되여있다.

AP011163.gbk는 게놈의 기본정보를 담고있는 화일로서 여기에는 게놈구성정보, 참고문헌정보, 게놈내유전자정보, 염기배렬정보가 모두 들어있다.(그림 1)

```
LOCUS       AP011163           2815241 bp    DNA     circular BCT 27-AUG-2009
DEFINITION  Acetobacter pasteurianus IFO 3283-01-42C DNA, complete genome.
ACCESSION   AP011163
VERSION     AP011163.1  GI:256650512
DBLINK      BioProject: PRJDA31141
KEYWORDS    .
SOURCE      Acetobacter pasteurianus IFO 3283-01-42C
  ORGANISM  Acetobacter pasteurianus IFO 3283-01-42C
            Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales;
            Acetobacteraceae; Acetobacter.
REFERENCE   1
  AUTHORS   Azuma,Y., Hosoyama,A., Matsutani,M., Furuya,N., Horikawa,H.,
            Harada,T., Hirakawa,H., Kuhara,S., Matsushita,K., Fujita,N. and
            Shirai,M.
  TITLE     Whole-genome analyses reveal genetic instability of Acetobacter
            pasteurianus
  JOURNAL   Unpublished
  REMARK    Publication_Status: Available-Online
REFERENCE   2  (bases 1 to 2815241)
  AUTHORS   Azuma,Y. and Hosoyama,A.
  CONSRTM   Acetobacter pasteurianus genome sequencing consortium
  TITLE     Direct Submission
  JOURNAL   Submitted (06-APR-2009) Contact:Yoshinao Azuma Yamaguchi University
            School of Medicine, Department of Microbiology and Immunology;
            Minami-Kogushi 1-1-1, Ube, Yamaguchi 755-8505, Japan URL
            :http://www.bio.nite.go.jp/dogan/Top
COMMENT     This work was done by Acetobacter pasteurianus genome sequencing
            consortium including Department of Microbiology and Immunology,
            Yamaguchi University School of Medicine, National Institute of
            Technology and Evaluation, Genetic Resources Technology, Faculty of
            Agriculture, Kyushu University, Department of Biological Chemistry,
            Faculty of Agriculture, Yamaguchi University and Mizkan Group
            Corporation Central Research Institute.
FEATURES             Location/Qualifiers
     source          1..2815241
                     /organism="Acetobacter pasteurianus IFO 3283-01-42C"
                     /mol_type="genomic DNA"
                     /strain="IFO 3283"
                     /sub_strain="IFO 3283-01-42C"
                     /db_xref="NBRC:105190"
                     /db_xref="taxon:634458"
                     /note="strain coidentity: IFO 3283 = NBRC 3283"
       gene          complement(388..813)
   … … …
ORIGIN
        1 actgcaggcg tcgagttcat tgaagagaaa ggtggtgggg ctggagttag gttgaggaaa
… … …
```

그림 1. AP011163.gbk의 일부

강조체는 식별기호를, 밑선친 부분은 부분식별기호를 의미한다.

매 행의 앞에는 그 행이 어떤 내용을 포함하고있는가를 나타내는 식별기호(그림에서 강조체부분)가 있는데 보다 세부적인 정보를 나타내는 경우에는 그 아래행들에 부분식별기호(그림에서 밑선친 부분)와 함께 해당한 정보를 주었다.

AP011163.rpt는 게놈의 등록정보를 포함하는 화일로서 종목록표를 작성하는데 리용하였다.(그림 2)

```
Accession: AP011163.1
GI: 256650512
DNA  length = 2815241
Taxname: Acetobacter pasteurianus IFO 3283-01-42C
Taxid: 634458
Genetic Code: 11
Protein count: 2562
CDS count: 2562
Pseudo CDS count: 0
RNA count: 66
Gene count: 2628
Pseudo gene count: 0
Others: 2628

Total: 5256
```

그림 2. AP011163.rpt의 내용

AP011163.gff는 게놈내의 유전자들을 위치별로 차례차례 렬거한 화일로서 그 내용이 AP011163.gbk와 류사하다.(그림 3)

```
AP011163.1 DDBJ   gene   388    813    .      -      .
    ID=gene0;Name=rusA;gbkey=Gene;gene=rusA;locus_tag=APA42C_00010;part=1%2F1
```

그림 3. AP011163.gff의 일부

그림 3은 AP011163.gff의 내용에서 대표적인 한행을 보여주고있다. 여기서 기본자료마당들은 Tab기호로 구분되여있다. 그림에서 AP011163.1은 해당 게놈의 등록번호를, DDBJ는 해당 유전자가 DDBJ에 등록되여있다는것을 의미하며 388과 813은 게놈내에서 시작과 끝위치를, －는 전사방향이 부의 방향이라는것을 의미한다. 마지막마당은 해당 위치에 놓인 유전자의 이름과 등록번호와 같은 간단한 정보를 담고있다.

gbk에서는 유전자를 단위로 하여 정보들이 기록되며 gff에서는 게놈의 시작위치로부터 마감위치까지 가면서 순차적으로 단편정보들이 기록된다. 하나의 유전자를 이루는 배렬들이 게놈의 여러곳에 분산되여 존재하는 경우 gbk에서는 유전자를 이루는 모든 단편정보들이 련속적으로 기록되지만 gff에서는 단편들이 위치별로 제각기 기록되므로 게놈지도를 작성하는데는 gbk보다 gff를 리용하는것이 유리하다.

AP011163.ptt(그림 4)와 AP011163.rnt(그림 5)는 각각 게놈내에 들어있는 단백질정보와 RNA정보만을 담고있는 화일로서 화일구성형식은 꼭같다. 우리는 이 화일들을 리용하여 단백질목록표와 RNA목록표를 작성하였다.

ptt와 rnt에서 한행은 각각 하나의 단백질정보, 하나의 RNA정보를 담고있으며 자료마당들은 Tab기호로 구분되여있다. 그림 4와 5에서 강조체는 해당 자료마당이름으로서 그 아래로 해당 단백질과 RNA의 자료들이 렬거되여있다.

AP011163.fna는 해당 게놈의 염기배렬정보를 FASTA형식으로 보관하고있다.

비루스게놈정보는 1개의 gbk에 들어있다. 비루스게놈정보들이 보통 수만건씩 1개의 GenBank형식의 화일에 들어있으며 그 형식은 세균의 gbk와 꼭같다.

Acetobacter pasteurianus IFO 3283-01-42C DNA, complete genome. - 1..2815241
2562 proteins

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
388..813	-	141	256650513	rusA	APA42C_00010	-	-	Holliday junction resolvase RusA
1206..2420	-	404	256650514	-	APA42C_00020	-	-	p h a g e integrase
2681..3160	-	159	256650515	-	APA42C_00030	-	-	hypothetical protein

… … …

그림 4. AP011163.ptt의 일부 내용
강조체는 해당 마당의 이름을 의미한다.

Acetobacter pasteurianus IFO 3283-01-42C DNA, complete genome. - 1..2815241
66 RNAs

Location	Strand	Length	PID	Gene	Synonym	Code	COG	Product
20268..20354	+	87	256650512	-	APA42C_00220	-	-	Leu tRNA
125774..125850	-	77	256650512	-	APA42C_01140	-	-	Trp Trna

… … …

그림 5. AP011163.rnt화일내용의 일부
강조체는 해당 마당의 이름을 의미한다.

우리는 Java를 리용하여 앞에서 언급한 형식의 게놈자료들로부터 MySQL자료기지를 구축하기 위한 프로그람을 작성하고 모든 세균 및 비루스게놈정보를 포함하는 세균 및 비루스게놈정보자료기지를 구축하였다.

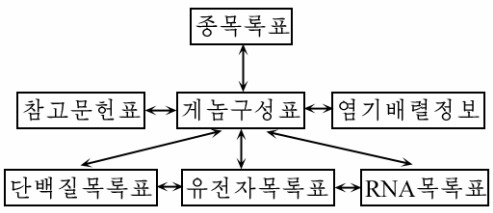

그림 6. 자료기지를 구성하고있는 표들사이 련관

자료기지를 종목록표, 게놈구성표, 참고문헌표, 유전자목록표, 단백질목록표, RNA목록표로 구성하였다. 염기배렬정보는 자료기지에 포함시키지 않고 화일로 따로 묶어 리용하였다. 자료기지를 구성하고있는 표들사이 련관을 그림 6에 보여주었다.

종목록표에는 자료기지로 구축된 세균, 비루스종목록을 보관하였으며 게놈구성표(표 2)와 참고문헌표(표 3)에는 해당 게놈의 기본구성정보들과 참고문헌정보들을 보관하였다.

표 2. 게놈특성표의 기본마당구조

마당이름	자료의미	자료의 례
GENOME_ID	자료기지내에서의 게놈ID	12
LOCUS	게놈간략정보(등록번호, 염기배렬 길이, 형태, 등록날자)	AP011163 2815241 bp DNA circular BCT 27-AUG-2009
ACCESSION	등록번호	AP011163
VERSION	등록번호.판번호	AP011163.1
GI	웅근수로 표현된 Genbank유일번호	256650512
DEFINITION	게놈명	*Acetobacter pasteurianus* IFO 3283-01-42C DNA, complete genome
DATE	NCBI등록날자	27-AUG-2009
TYPE	게놈DNA형태	circular BCT
COMMENT	주석정보	This work was done by *Acetobacter pasteurianus* genome sequencing … …
SIZE	게놈크기	2815241
DBLINK	다른 자료기지와의 련결정보	BioProject: PRJDA31141
KEYWORDS	실마리어	Complete genome… …
SOURCE	해당 게놈이 분리된 생물종명	*Acetobacter pasteurianus* IFO 3283-01-42C

표 3. 참고문헌표의 기본마당구조

마당이름	자료의미	자료의 례
REFERENCE_ID	자료기지내에서의 참고문헌ID	6061
TITLE	제목	Complete genome sequence determination of a *Macrococcus caseolyticus* strain JSCS5402 reflecting the ancestral genome of the human pathogenic staphylococci
JOURNAL	잡지이름	J. Bacteriol. (2008) In press
AUTHORS	저자	Baba T., … …
PUBMED	의학정보자료기지련결정보	19074389
REMARK	기타 문헌관련정보	Publication Status: Available-Online prior to print
GENOME_ID	자료기지내에서의 게놈ID	2750

모든 표에는 GENOME_ID라는 자료마당이 있는데 이 자료마당을 리용하여 모든 표들이 게놈특성표와 련결되여있다.

유전자목록표는 gff와 gbk에 기초하여 2개로 작성하였다.(표 4와 5)

표 4. gff에 기초하여 만든 유전자목록표의 마당구조

마당이름	자료의미	자료의 례
TYPE	등록된 자료기지이름	EMBL
REGION	유전자류형	CDS
START	시작위치	241
END	끝위치	495
STRAND	전사방향	+
CONTENT	유전자정보	Name=CCB80881.1; … …
GENOME_ID	자료기지내에서의 게놈ID	157

표 5. gbk에 기초하여 만든 유전자목록표의 마당구조

마당이름	자료의미	자료의 례
REGION	유전자류형	CDS
LOCATION	게놈내에서 유전자위치	complement(217..633)
LOCUS	NCBI의 유전자등록번호	HP_0001
CONTENT	유전자정보	/note="hypothetical protein; identified by GeneMark; …
GENOME_ID	자료기지내에서의 게놈ID	157

gff로 만든 유전자목록표는 게놈지도를 그리는데 리용되며 gbk로 만든 표는 해당 유전자정보를 구체적으로 열람하려고 할 때 리용된다.

이밖에도 우리는 단백질목록표와 RNA목록표도 만들었는데 이것들을 유전자목록표와 련결하여 해당 유전자정보의 열람이나 개별적인 단백질, RNA의 검색에도 리용되게 하였다.

또한 우리는 게놈염기배렬정보들을 모두 합하여 FASTA형식으로 만들어 상동성검색의 입력화일로 리용할수 있게 하였다.

결과적으로 2 854종에 달하는 세균의 4 852개 게놈자료, 102 229종에 달하는 비루스의 1 433 313개 게놈자료가 자료기지로 구축되였으며 여기서 자료기지내의 염기배렬자료용량은 세균인 경우 8.32GB, 비루스인 경우 2.11GB였다.

2) 망에서 리용할수 있는 세균 및 비루스게놈정보검색 및 열람체계의 개발

우리는 구축한 세균 및 비루스게놈자료기지에 기초하여 세균 및 비루스게놈정보검색 및 열람체계를 개발하였다.

세균 및 비루스게놈정보검색 및 열람체계는 크게 게놈정보열람, BLAST상동성검색, 세균유전자검색부분으로 구성하였다. (그림 7) 여기서 열람체계는 게놈구성정보열람, 게놈지도열람, 유전자열람부분으로, 검색체계는 게놈구성정보(NCBI등록번호, 게놈이름)에 의한 게놈검색과 BLAST배렬상동성에 의한 게놈검색, 세균유전자검색부분으로 구성하였다.

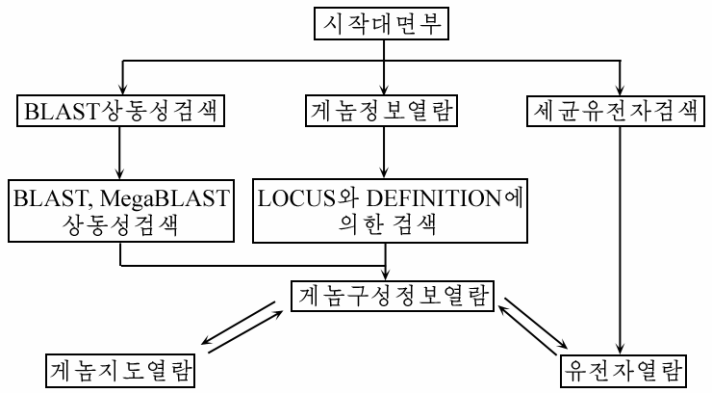

그림 7. 세균 및 비루스게놈정보검색 및 열람체계 구성도

게놈구성정보열람 게놈구성정보열람창에서는 게놈구성표와 참고문헌표에 보관된 게놈관련정보들을 열람할수 있게 하였다. 비루스게놈인 경우 게놈내에 들어있는 유전자정보량이 적은것으로 하여 유전자목록정보와 염기배렬정보도 같이 열람할수 있게 하였다.

게놈지도열람 게놈지도열람창에서는 게놈내에서 유전자들의 분포상태를 지도로 볼수 있게 하였다. 세균게놈지도는 유전자와 CDS, RNA부분으로 나누었는데 전사방향까지 고려하여 총 6개 층으로 구성하였으며 유전자들은 종류와 전사방향에 따라 해당 층과 위치에 현시되게 하였다.

비루스게놈에서 유전자들은 같은 자리의 DNA배렬을 반복리용하는 빈도가 매우 높으므로 우리는 비루스게놈지도를 10여개의 층으로 만들어 서로 반복리용되는 자리의 유전자들이 겹치여 현시되지 않게 하였으며 전사방향에 따라 그 색갈을 달리하여 구분해주었다. 게놈지도열람창에서는 해당 위치의 염기배렬도 같이 볼수 있게 하였다. 오른쪽 웃부분에 현시설정창을 배치하여 열람하려는 게놈부분을 선택할수 있게 하였다.

유전자자료열람 유전자자료열람창에서 게놈내에 들어있는 유전자목록을 표로 현시해주었으며 필요한 유전자에 대하여 구체적인 정보들을 열람할수 있게 하였다. 유전자목록에서는 필요한 항목에 한하여 검색도 진행할수 있게 하였다.

배렬상동성검색 ubuntu봉사기체계우에서 NCBI가 제공하는 wwwBLAST를 개조하여 망을 통한 배렬상동성검색을 할수 있게 하였다. wwwBLAST에서 제공하는 기능들가운데서 BLAST와 MegaBLAST만을 리용하였으며 BLAST자료기지는 비루스와 세균게놈의 염기배렬정보들을 FASTA형식의 자료로 형식화하고 formatdb프로그람을 리용하여 만들었다.

배렬상동성검색결과물들은 모두 자료기지화된 게놈자료들로서 게놈구성정보열람창, 지도열람창, 유전자열람창에서 열람할수 있다.

세균유전자검색 세균유전자검색창에서는 세균게놈자료기지내에 들어있는 유전자들에 대한 검색을 진행할수 있게 하였다. 검색하려는 세균종들을 선택하여 해당 종에 한해서만 검색을 진행할수 있게도 하였다. 검색된 결과는 유전자자료열람창에서 열람할수 있으며 해당 유전자를 가지고 있는 세균게놈정보도 게놈구성정보열람창과 게놈지도열람창을 통하여 열람할수 있게 하였다.

맺 는 말

프로그람을 리용하여 게놈자료들을 해석하여 세균, 비루스게놈자료기지, BLAST자료기지를 구축하였다.

망을 통하여 세균, 비루스게놈자료들을 여러가지 방법으로 검색하여 열람할수 있는 게놈정보열람체계를 개발하였다.

참 고 문 헌

[1] R. Leinonen et al.; Bioinformatics, **22**, **10**, 1284, 2006.

[2] www.ncbi.nlm.nih.gov\PMGifs\Genomes\mier.html, **4**, 2016.

[3] N. F. Alikhan et al.; BMC Genomics, **12**, 402, 2011.

주체106(2017)년 1월 5일 원고접수

Development of Browsing Software of Bacterial and Viral Genome Information

Hwang Jin Hyok, Ri Sok Jun

Recently many species of bacterial and viral genome information is analyzed and widely distributed through the internet, and thus this information is variously used in biological research.

We collected the text type of genome data of bacteria and virus on the internet and developed the effective algorithm to extract the useful data to construct genome and BLAST databases which are related each other, and developed the browsing software of bacterial and viral genome information.

Key words: bacteria, virus, genome, BLAST, database