Genbank 분석을 통한 NDSL 연계 서비스 모형 설계 연구

안부영*, 이정훈**, 김대환*, 신용주*, 최선희*, 신진섭*
*한국과학기술정보연구원 콘텐츠융합팀
**(주)킨폭스
e-mail:*{ahnyoung, dhkim, yjshin, sunny.choi, js.shin}@kisti.re.kr,
**exturbo@paran.com

A Study on Design of NDSL Linked Service Model by Analysis of Genbank

Bu-Young Ahn*, Jung-Hun Lee**, Dea-Hwan Kim*, Yong-Ju Shin*,
Seon-Heui Choi*, Jin-Seob Shin*
*Content Convergence Team, Korea Institute of Science & Technology Information
**Kinfox Corporation

요 약

최근 들어 분자생물학의 급속한 발전과 2001년 인간유전체사업의 완료로 인해 전세계적으로 엄청난 양의 유전정보가 공개되었다. 유전자 서열정보는 그 양이 방대하고 다양하기에 데이터베이스 구축 및 분석을 위하여 고성능 컴퓨터 및 정보기술 기법이 필요하다. 그래서 컴퓨터를 활용하여 생물학적 데이터를 수집, 관리, 저장, 평가, 분석하는 연구분야인 생명정보학(바이오인포매틱스)이라는 학문이 지속적으로 발전하고 있다. 이런 생명정보학 발전에 발맞추어 한국과학기술정보연구원(KISTI)에서는 정보기술을 기반으로 한 생명정보 인프라를 구축하여 생명과학 연구자들에게 제공하고 있다. 본 논문에서는 생명정보 데이터베이스중에서 연구자들이 가장 많이 이용하는 유전자 데이터베이스인 Genbank를 활용 및 분석하여 KISTI에서 운영하는 학술논문 제공 사이트인 NDSL(http://scholar.ndsl.kr)과 연계 가능한 서비스 모델을 개발하기 위하여 1) NCBI FTP 사이트에서 Genbank 데이터를 수집하고, 2) Genbank 텍스트 파일을 유전자 기본정보와 참고 데이터베이스로 재구축하며, 3) Genbank refrence 필드에서 논문 및 특허 정보 추출을 통한 새로운 테이블을 생성하여 NDSL과 연계 가능한 서비스 모델을 제안하였다.

1. 서론

2001년 인간유전체사업(Human Genome Project)의 완료로 인해 전세계적으로 엄청난 양의 유전정보가 공개되어 인간 유전자 지도가 완성되었고, 인간의 유전자는 어떤 화학적 염기배열로 구성된 것이 밝혀졌다. 게놈(genome)이란 유전자(gene)와 염색체(chromosome)의 합성어이다.

유전자 서열정보는 그 양이 방대하고 다양하기에 컴퓨터를 활용한 분석 및 이를 활용 가능한 정보기술이 필요하다. 그래서 컴퓨터를 활용하여 생물학적 데이터를 수집, 관리, 저장, 평가, 분석하는 연구분야인 생명정보학(바이오인포매틱스)이 지속적으로 발전하고 있다. 생명정보학을 간단하게 설명하자면 생물학실험실을 컴퓨터로 옮겨 놓은 것이라 말할 수 있다.

본 논문에서는 생명정보 데이터베이스중에서 전세계적으로 연구자들이 가장 많이 이용하는 유전자 데이터베이스인 Genbank를 대상으로 Genbank의 reference 필드를 분석하여 논문 정보(논문제목, 저자, 수록처 등) 및 특허정보를 추출하여 KISTI에서 운영하는 학술논문 제공 사이트인 NDSL(http://scholar.ndsl.kr)과 연계 가능한 서비스 모델을 제안하고자 한다.

2. Genbank 소개 및 분석

2.1 Genbank 소개

인간은 약 100조개의 세포로 구성되어 있으며, 세포내에는 세포핵이 존재한다. 세포는 23쌍의 염색체로, 23쌍의 염색체는 31억개의 염기쌍으로 구성되어 있으며, 염기는 시토신(C), 구아닌(T), 아데닌(A), 티민(T)으로 구성되어 있다. 이렇게 규명된 유전자 염기서열을 데이터베이스로 구축하여 인간의 질병연구 및 치료에 활용하고 있다.

이와 같은 유전자 데이터베이스중에서 전세계적으로 가장 많이 사용되는 것은 미국 국립보건원(NIH, National Institutes of Health)의 국립생물공학정보센터(NCBI, National Center for Biotechnology Information)에서 운영하는 Genbank이다. Genbank는 염기서열정보 데이터베이스로 세계 각지에서 연구자들이 각자 등록한 서열 데이터를 다양한 각도의 분석결과와 함께 제공한다. GenBank와 실시간으로 데이터 미러를 하는 기관은 유럽의 유럽분자생물학실험실(EMBL, European Molecular Biology Laboratory)과 일본의 DNA데이터뱅크(DDBJ, DNA Data Bank of Japan)가 있다[1]. 2008년 8월 현재 release 167 기준으로 약 9천 2백만건의 염기 서열을 제공하고 있다.

NCBI는 또한 생물, 의학분야 최대 문헌정보서비스인 Pubmed를 운영하고 있기에 (그림 1)에서 보는 바와 같이 Genbank reference 필드에 Pubmed id를 링크하여 연계 서비스를 제공하고 있다. 그러나 Pubmed에 등재되지 않은 논문은 링크되어 있지 않아 Pubmed 이외의 논문을 필요로 하는 이용자들에게 불편함을 주고 있다.

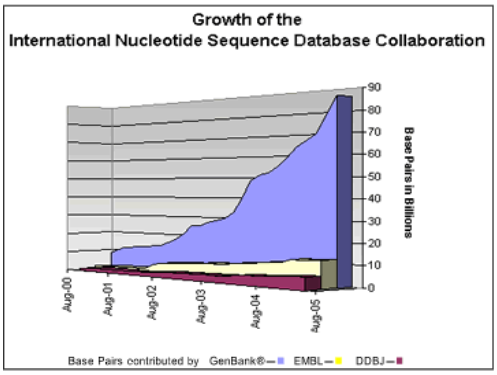
```
LOCUS       XELSRCC2                 115 bp    mRNA    linear   VRT
28-APR-1993
DEFINITION  X.laevis Rous sarcoma virus transforming protein mRNA, 3' end.
ACCESSION   M30860
VERSION     M30860.1  GI:214812
KEYWORDS    transforming protein.
SEGMENT     2 of 2

SOURCE      Xenopus laevis (African clawed frog)
  ORGANISM  Xenopus laevis
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi;
            Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea;
            Pipidae;
            Xenopodinae; Xenopus; Xenopus.
REFERENCE   1  (bases 1 to 115)
  AUTHORS   Steele,R.E.
  TITLE     Two divergent cellular src genes are expressed in Xenopus
laevis
  JOURNAL   Nucleic Acids Res. 13 (5), 1747-1761 (1985)
  MEDLINE   85215578
  PUBMED    2987836

COMMENT     Original source text: X.laevis (female) erythrocyte, cDNA to
mRNA.
FEATURES             Location/Qualifiers
     source          1..115
                     /organism="Xenopus laevis"
                     /mol_type="mRNA"
                     /db_xref="taxon:8355"
     CDS<1..69
                     /note="transforming protein (src)"
                     /codon_start=1
                     /protein_id="AAA49965.1"
                     /db_xref="GI:214815"
                     /translation="LQAFLEDYFTATEPQYQPGDNL"
ORIGIN      Undetermined number of bp after segment 1.
        1 ctgcaggcgt tcctggagga ctattttaca gctaccgaac cgcagtacca
gcctggggac
       61 aacctttagg cttcgctcat aatcaagaga catgtatagg actcttagga aacag
//
```

(그림 1) GenBank 데이터 구성[2]

Genbank에 등록되는 유전자서열은 급속하게 증가하고 있으며 (그림 2)는 NCBI, EMBL, DDBJ에서 등록되는 유전자 서열 증가 추세를 볼 수 있는 그래프이다.



(그림 2) Genbank 데이터 증가현황[3]

2.2 Genbank 분석

NCBI, EMBL, DDBJ에서는 Genbank를 무상으로 다운로드할 수 있도록 FTP 사이트를 운영하고 있다. 본 논문에서는 Genbank release 163 기준으로 약 8천 4백만건의 데이터를 분석하여 필요한 필드를 추출하였다.

2.2.1 reference-논문 필드 분석

Genbank 필드중에서 reference 필드는 <표 1>과 같은 항목으로 구성되어 있으며 reference는 유전자정보 한 개당 N개까지 기술이 가능하다. 전체 데이터의 reference 필드를 분석해 본 결과 약 84백만건의 유전자정보, 1억건정도의 reference 건수를 확인할 수 있었다. 본 분석결과로 보면 유전자정보 1건당 1.2개의 reference가 기술되어 있다는 것을 알 수 있다.

<표 1> Genbank refrence 필드 구성

필드명	데이터 기술 예
REFERENCE	1 (bases 1 to 399)
AUTHORS	Belshaw,R., Fitton,M., Herniou,E., Gimeno,C. and Quicke,D.L.J.
TITLE	A phylogenetic reconstruction of the Ichneumonoidea (Hymenoptera) based on the D2 variable region of 28S ribosomal RNA
JOURNAL	Syst. Entomol. 23, 109-123 (1998)
MEDLINE	85215578
PUBMED	2987836

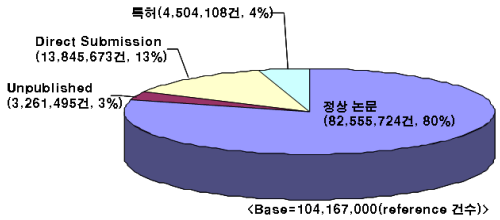
Genbank 데이터의 reference 필드를 추출하여 유형을 분석한 결과 <표 2>와 같이 정상적으로 필드가 기술된 경우, Unpublished인 경우, Direct submission인 경우, Patent인 경우 등 4가지 유형으로 나타났다.

<표 2> refrence 필드 유형

유형	필드명	기술 내용
정상	AUTHORS	Harano,Y., Suzuki,I., Maeda,S., Kaneko,T., Tabata,S. and Omata,T.
	TITLE	Identification and nitrogen regulation of the cyanase gene from the cyanobacteria Synechocystis sp. strain PCC 6803 and Synechococcus sp. strain PCC 7942
	JOURNAL	J. Bacteriol. 179 (18), 5744-5750 (1997) 9294430
Unpublished	AUTHORS	Chen,W. and He,W.B.
	TITLE	Nucleotide Sequence and Characteristics of beta-amylase Gene from Bacillus firmus
	JOURNAL	Unpublished
Direct Submission	AUTHORS	Iwabuchi,T.
	TITLE	Direct Submission
	JOURNAL	Submitted (25-DEC-1996) Tokuro Iwabuchi, Shiseido Research Center, Pharmaco Science Laboratories: 1050 Nippa, Kouhoku-ku, Yokohama, Kanagawa 223, Japan (E-mail:PEH01461@niftyserve.or.jp, Tel:+81-45-542-1337, Fax:+81-45-545-5931)
Patent	AUTHORS	Koizumi,S., Yonetani,Y. and Teshiba,S.
	TITLE	Process for producing riboflavin
	JOURNAL	Patent: US5589355-A 31-DEC-1996;

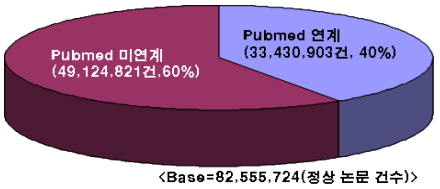
Genbank는 텍스트파일로 FTP 사이트를 통하여 130여개의 압축된 파일로 제공된다. 그래서 연계를 위하여 텍스

트파일의 압축을 풀어 유전자 기본정보와 reference 정보를 추출하여 MySQL 데이터베이스로 변환하였다. 변환작업 결과 유전자정보 건수는 84,112,248건, reference 건수는 104,167,000건으로 나타났으며 reference 유형별 데이터 분포는 (그림 3)과 같다.



(그림 3) reference 유형별 분포 현황

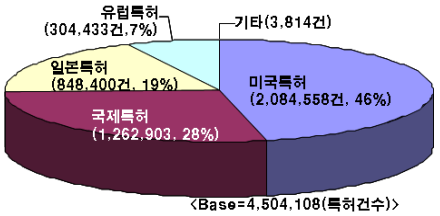
(그림 3)의 정상 논문 82,555,724건중에서 Pubmed id를 가지고 있는 논문은 33,430,903건(40%)이고, Pubmed id를 가지고 있지 않은 논문은 49,124,821건(60%)이었다. 즉 Genbank 유전자정보의 reference-논문정보-중 60%정도가 Pubmed와 연계되어 있지 않다는 결과를 나타낸다.



(그림 4) Pubmed 연계 논문 현황

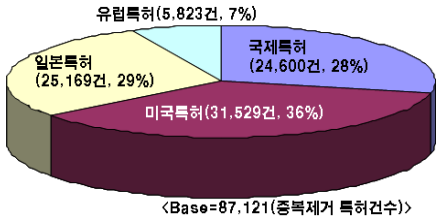
2.2.2 reference-특허 필드 분석

<표 2>의 reference 필드 유형 중 특허정보가 기술된 경우는 약 4백 5십만건정도였다. 각국별 공개 또는 등록된 특허 건수를 산출한 결과는 (그림 5)와 같다.



(그림 5) 국가별 특허 현황

(그림 5)의 국가별 특허 현황에서 정상적으로 매핑이 가능한 데이터를 추출하기 위하여 중복을 제거한 후 각국별로 특허건수를 산정한 결과는 (그림 6)과 같다. 두 개의 결과에서 볼 수 있듯이 미국이 가장 많은 특허를 보유하고 있으며, 국제특허, 일본, 유럽이 그 뒤를 따르고 있었으며, 대한민국 특허는 4건이 등록되어 있었다.



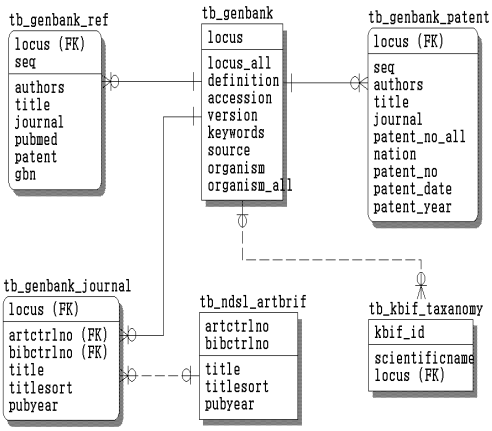
(그림 6) 중복제거 후 특허 분포현황

3. 연계 모형 설계

2장에서는 Genbank 데이터중에서 reference 필드를 분석하여 그 결과를 알아 보았다. 3장에서는 2장의 분석 결과를 바탕으로 NDSL과 연계하기 위한 매핑 테이블 및 KISTI CCBB 웹사이트(http://www.ccbb.re.kr)에서 서비스중인 Genbank 데이터베이스를 중심으로 연계 모형을 설계하였다.

3.1 테이블 설계

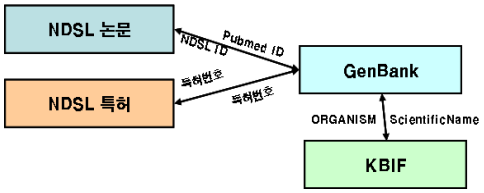
본 연계 모형에서 사용될 데이터베이스 테이블은 (그림 7)의 ERD(Entity Relationship Diagram)에서 보는 것과 같이 Genbank에서 추출한 유전자 기본정보를 저장할 테이블(tb_genbank), reference 필드를 저장할 테이블(tb_genbank_ref), reference 필드의 특허정보를 저장할 테이블(tb_genbank_patent), reference 필드의 논문정보를 저장할 테이블(tb_genbank_journal, tb_ndsl_artbrif), organism 필드에서 추출한 생물다양성 종정보를 저장할 테이블(tb_genbank_taxonomy) 등 6개이다. (그림 7)과 같이 설계된 6개의 테이블을 활용하여 시스템간 연계 및 화면을 설계하였다.



(그림 7) ERD(매핑을 위한 테이블)

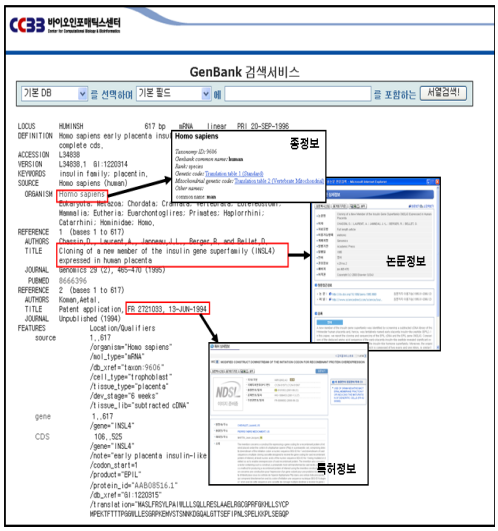
3.2 시스템 연계 및 검색 화면 설계

시스템 연계를 위하여 첫번째로 NDSL 논문과 GenBank reference 필드의 title(논문명)간의 Mapping을 통한 NDSL ID 추출 및 연계를 수행하려 한다. 두번째로 NDSL 특허와 GenBank의 reference 필드중 patent 부분에서 추출한 특허번호를 정제하여 정제된 특허번호와 NDSL 특허번호를 연계하여 특허정보의 서지 및 원문정보를 검색 가능하게 하려고 한다. 세번째로 KBIF의 생물다양성정보에서 검색되는 학명(Scientific Name)과 GenBank source 필드 중 organism 부분의 종(taxonomy) 정보을 연계하려고 한다. (그림 8)은 각 시스템간 연계를 가능하게 하는 주요 키를 나타내고 있다.



(그림 8) 시스템간 연계 구성도

Genbank 데이터 필드는 Locus(유전자 위치), Definition(생물학적 특성), Accession(고유 식별자), Keywords(서열의 핵심어), Source(생물체의 일반명/학명), Reference(인용된 논문사항), Comment(주석 및 해설), Features(생물학적 특징정보), Origin(서열의 원 소스) 등 크게 9개로 구성되어 있다. 이중에서 본 연계 모형에 매핑될 필드는 Source 필드중에서 organism 부분, Reference 필드중에서 논문정보와 특허정보 부분이다.



(그림 9) 콘텐트 연계 서비스 설계 화면

(그림 9)는 특정 필드를 추출하여 매핑 테이블을 작성한 후 CCBB 웹사이트에서 검색한 유전자 정보에 NDSL을 링크하여 연계 서비스되는 화면을 보여주는 것이다.

4. 결론 및 향후 연구방향

지금까지 Genbank 데이터베이스를 활용하여 각 필드를 분석하여 그 결과를 산출해 보았으며, 산출된 결과를 기본으로 연계를 위한 매핑 테이블을 설계하였다. 향후에는 Genbank의 organism 필드를 분석한 후 KISTI에서 서비스하고 있는 약 110만건의 생물다양성 데이터의 종(taxonomy) 정보와 연계 가능하도록 매핑 테이블을 설계하고, 설계된 내용을 바탕으로 2008년내로 코딩을 완료하여 프로토타입을 완성할 계획이다.

본 논문에서 제안한 프로토타입이 시스템으로 구현되어 서비스된다면 기존 GenBank에서 제공되지 않는 pubmed id 미보유 논문중 상당수의 논문을 NDSL 연계로 제공 가능해 질것이며, 미국 특허를 위주로 서비스되고 있는 Genbank와 KISTI에서 보유하고 있는 유럽 및 일본 특허정보와의 연계 또한 가능해지고, 국내 생물다양성정보와의 연계로 인하여 생명과학 연구자들에게 좀 더 유용한 고부가가치 서비스를 제공할 수 있을 것으로 기대된다.

참고문헌

[1] 안부영, 한정민, 한건, 이상호, “생명정보 연계검색 인터페이스 설계에 관한 연구”, 제29회 한국정보처리학회 춘계학술발표대회 논문집 15(1): 407-409, 2008.5.

[2] 안부영, 오충식 “생명정보 콘텐트 업데이트 가이드 v. 2.0”, ISBN 978-89-6211-245-0, 한국과학기술정보연구원, 2008. 8.

[3] NCBI(Genbank) FTP 사이트, ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

[4] KISTI 바이오인포매틱스 웹사이트, http://www.ccbb.re.kr

[5] KISTI 과학기술정보 통합서비스 웹사이트, http://www.ndsl.kr