Genbank 분석을 통한 이종의 콘텐츠 연계 방안 설계

Design of Heterogeneous Content Linkage Method by Analyzing Genbank

안부영*, 이명선*, 김지영**, 오충식***
한국과학기술정보연구원 슈퍼컴퓨팅본부*, 정보유통본부**, 정보화전략팀***

Bu Young Ahn(ahnyoung@kisti.re.kr)*, Myung Sun Lee(mslee@kisti.re.kr)*,
Ji Young Kim(yes@kisti.re.kr)**, Chung Shick Oh(ocs@kisti.re.kr)***

요약

유전자 서열정보는 그 양이 방대하고 다양하기에 DB 구축 및 분석을 위하여 고성능 컴퓨터 및 정보기술 기법이 필요하다. 그래서 컴퓨터를 활용하여 생물학적 데이터를 수집, 관리, 저장, 평가, 분석하는 연구분야인 생명정보학이라는 학문이 지속적으로 발전하고 있다. 이런 생명정보학 발전에 발맞추어 한국과학기술정보연구원(KISTI)에서는 정보기술 기반 생명정보 인프라를 구축하여 생명과학 연구자들에게 제공하고 있다. 본 논문에서는 생명정보 DB 중에서 전세계 연구자들이 가장 많이 이용하는 유전자 DB인 Genbank의 reference 필드를 분석하여 한국과학기술정보연구원(KISTI)의 과학기술정보 통합서비스인 NDSL (http://NDSL.kr)과의 연계 방안을 제안하고자 한다. 이를 위하여 NCBI FTP 사이트에서 Genbank 데이터를 수집하여 Genbank 텍스트 파일을 유전자 기본정보와 참고정보로 나누어 DB로 재구축하였으며 Genbank reference 필드에서 논문 및 특허 정보 추출을 통한 새로운 테이블을 생성하였고, KISTI의 논문 DB (http://scholar.ndsl.kr), 특허 DB (http://patent.ndsl.kr)와의 연계 방안을 제시하였다.

■ 중심어 : | 콘텐츠 연계 | 유전자 데이터베이스 | 논문 콘텐츠 | 특허 콘텐츠 | Genbank |

Abstract

As information on gene sequences is not only diverse but also extremely huge in volume, high-performance computer and information technology techniques are required to build and analyze gene sequence databases. This has given rise to the discipline of bioinformatics, a field of research where computers are utilized to collect, to manage, to save, to evaluate, and to analyze biological data. In line with such continued development in bioinformatics, the Korea Institute of Science and Technology Information (KISTI) has built an infrastructure for the biological information, based on the information technology, and provided the information for researchers of bioscience. This paper analyzes the reference fields of Genbank, the most frequently used gene database by the global researchers among the life information databases, and proposes the interface method to NDSL which is the science and technology information integrated service provided by KISTI. For these, after collecting Genbank data from NCBI FTP site, we rebuilt the database by separating Genbank text files into the basic gene data and the reference data. So new tables are generated through extracting the paper and patent information from Genbank reference fields. Then we suggest the method of connection with the paper DB and the patent DB operated by KISTI.

■ keyword : | Content Linkage | Gene Database | Literature Content | Patent Content | Genbank |

* 본 논문은 한국콘텐츠학회 ICCC2009 국제학술대회 우수논문입니다.
접수번호 : #100421-002
접수일자 : 2010년 04월 21일
심사완료일 : 2010년 06월 09일
교신저자 : 김지영, email : yes@kisti.re.kr

I. 서 론

2001년 인간유전체사업(Human Genome Project)의 완료로 인해 전 세계적으로 엄청난 양의 유전정보가 공개되어 인간 유전자 지도가 완성되었고, 인간의 유전자는 어떤 화학적 염기서열로 구성된 것이 밝혀졌다. 유전체(genome)란 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(Bioinformatics, 바이오인포매틱스)이 지속적으로 발전하고 있다. 생명정보학을 간단하게 설명하자면 생물학 실험실을 컴퓨터로 옮겨 놓은 것이라 말할 수 있다[1]. 본 논문에서는 생명정보 데이터베이스 중에서 전 세계적으로 연구자들이 가장 많이 이용하는 유전자 데이터베이스인 Genbank를 대상으로 Genbank의 reference 필드에서 논문정보(논문제목, 저자, 수록처), 특허정보(특허명칭, 특허번호)를 분석 및 추출하여 KISTI에서 구축하여 운영하는 과학기술정보통합서비스인 NDSL(http://NDSL.kr)과의 연계 방안을 제안하고자 한다.

II. Genbank 개요 및 분석

1. Genbank 개요

인간은 약 100조개의 세포로 구성되어 있으며, 세포 내에는 세포핵이 존재한다. 세포는 23쌍의 염색체로, 23쌍의 염색체는 31억 개의 염기쌍으로 구성되어 있으며, 염기는 시토신(C), 구아닌(G), 아데닌(A), 티민(T)으로 구성되어 있다. 이렇게 규명된 유전자 염기서열을 데이터베이스로 구축하여 인간의 질병연구 및 치료에 활용하고 있다. 이와 같은 유전자 데이터베이스 중에서 전 세계적으로 가장 많이 사용되는 것은 미국 국립보건원(NIH, National Institutes of Health)의 국립생물공학정보센터(NCBI, National Center for Biotechnology Information)에서 운영하는 Genbank이다. Genbank는 염기서열 데이터베이스로 세계 각지에서 연구자들이 등록한 서열 데이터를 다양한 각도의 분석 결과와 함께 제공한다[3]. 또한 NCBI는 생물, 의학분야 최대 문헌정보서비스인 Pubmed를 운영하고 있기에 Genbank reference 필드에 Pubmed id를 링크하는 서비스를 제공하고 있다. 그러나 Pubmed에 등재되지 않은 논문은 링크되어 있지 않아 Pubmed 이외의 논문을 필요로 하는 이용자에게 불편함을 주고 있다.

2. Genbank 분석

NCBI, EMBL, DDBJ에서는 Genbank를 무상으로 다운로드 할 수 있도록 FTP 사이트를 운영하고 있다. 본 논문에서는 2007년 12월 현재 Genbank release 163 기준 약 8천 4백만 건의 데이터를 다운로드 받아 분석하여 필요한 필드를 추출하였다. 참고로 Genbank 원본 파일의 개수는 1,380개이며 압축 해제 시 파일당 약 250MB, 약 400만 라인 정도였다.

2.1 reference 논문 필드 분석

Genbank 필드 중에서 reference 필드는 [표 1]과 같은 항목으로 구성되어 있으며 reference는 유전자정보 한 개당 N개까지 기술이 가능하다.

표 1. Genbank reference 필드 구성[4]

필드명	데이터 기술 내용 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

전체 데이터의 reference 필드를 분석해 본 결과 약 8천 4백만 건의 유전자정보, 1억건 정도의 reference 건수를 확인할 수 있었다. 이 결과로 유전자정보 1건당 1.2개의 reference가 기술되어 있다는 것을 알 수 있다.

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. Genbank reference 필드 4가지 유형

유형	필드명	데이터 기술 내용 예
정상	AUTHORS	Harano,Y., Suzuki,I.,Maeda,S.,Kaneko, T.,Tabata,S.andOmata,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;

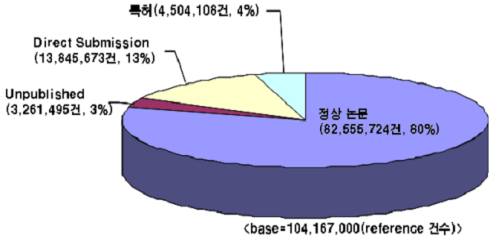

그림 1. reference 유형별 분포 현황

Genbank DB는 FTP 사이트를 통하여 다수의 압축 텍스트 파일로 제공된다. 콘텐츠 연계를 위하여 텍스트 파일의 압축을 풀어 유전자 기본정보와 reference 정보를 추출하여 MySQL 데이터베이스로 변환하였다. 변환 작업 결과 유전자정보 건수는 84,112,248건, reference 건수는 104,167,000건으로 나타났으며 reference 유형별 데이터 분포는 [그림 1]과 같다.

[그림 1]의 정상논문 82,555,724건 중에서 Pubmed id를 가지고 있는 논문은 33,430,903건(40%)이고, Pubmed id를 가지고 있지 않은 논문은 49,124,821건(60%)이었다. [그림 2]에서 보는 바와 같이 Genbank 유전자정보의 reference가 논문인 경우 60%정도가 Pubmed와 연계되어 있지 않다는 것을 알 수 있었다.

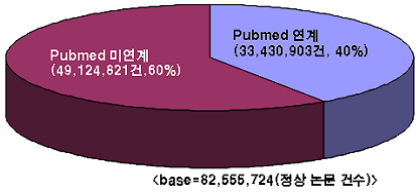

그림 2. Pubmed 연계 논문 현황

2.2 reference 특허 필드 분석

[표 2]의 reference 필드 유형 중 특허정보가 기술된 경우는 약 4백 5십만 건 정도였다. 각국별 공개 또는 등록된 특허 건수를 산출한 결과는 [그림 3]과 같다.

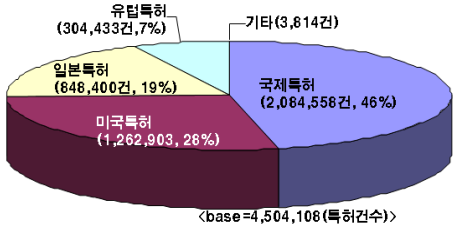

그림 3. 국가별 특허 현황

[그림 3]의 국가별 특허 현황에서 정상적으로 맵핑이 가능한 데이터를 추출하기 위하여 중복을 제거한 후 각 국별로 특허건수를 산정한 결과는 [그림 4]와 같다. 두 개의 결과에서 볼 수 있듯이 국제특허의 건수가 가장 많았지만 중복을 제거한 후에 비교해 본 결과 미국이

가장 많은 특허를 보유하고 있으며, 국제, 일본, 유럽이 그 뒤를 따르고 있었고, 대한민국 특허는 4건이 등록되어 있었다.

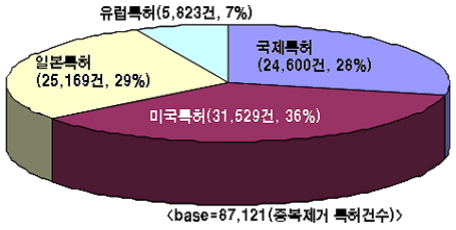

그림 4. 중복제거 후 특허 현황

III. Genbank와 NDSL 이종의 콘텐츠 연계 방안

1. NDSL 논문 콘텐츠 연계 방안

전체 논문 82,555,724건 중 Pubmed id 보유 33,430,903 건을 제외한 pubmed id 미 보유 논문에 대해 NDSL 논문과의 연계를 시도하려고 한다. 이를 위한 Pubmed id 미 보유 논문의 중복 제거 작업을 수행하면 서버의 메모리(4GB)를 초과하게 된다. 그래서 locus의 처음 2자리를 잘라서 영역을 나누고 200~400만 건 단위로 중복을 제거한 후 tb_genbank_journal_distinct에 로딩 하였으며, tb_genbank_journal_distinct의 데이터를 다시 중복 제거하여 tb_genbank_journal_distinct1에 입력하는 절차를 수행하였다.

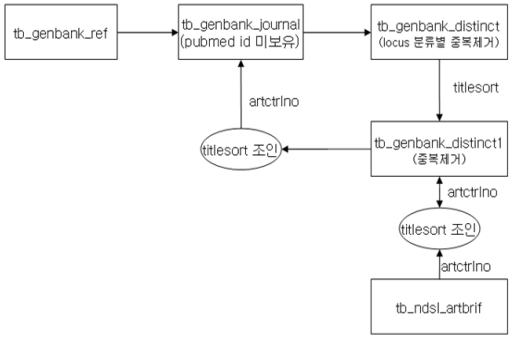

그림 5. Genbank reference 논문필드 정제 절차

논문 제목의 가공을 통한 NDSL ID(문헌번호)의 연계를 위하여 NDSL과 Genbank 논문의 제목에서 특수문자와 공백을 제거하고, 제목을 소문자로 변환한 후, NDSL논문 전체와 Genbank Pubmed id 미 보유 논문에 대해서 동일 제목을 가진 NDSL ID를 추출한 결과 아래와 같은 결과를 얻었다.

- 중복 제거 전 논문건수: 49,124,821건
- 중복 제거 후 논문건수: 68,515건
- 중복 제거 논문건수 중 NDSL 제목과 일치하는 건수: 19,370건
- 중복 제거 논문건수 중 제목과 발행년도 일치하는 건수: 16,921건
- Pubmed id 중복 제거 건수: 175,989건

논문 제목을 통한 연계시의 문제점은 제목이 동일한 다른 논문이 존재할 가능성이 있다는 것인데 이를 해결하기 위하여 논문 제목 이외에 저널 제목 및 발행연도를 추가로 비교하여 분석하였다. NDSL 저널 발행연도와 Genbank 저널 정보의 발행연도를 비교하여 검증한 결과 전체 19,370건의 중복제거 논문 중 16,921건의 데이터가 발행연도가 일치한다는 결과를 얻었다. 분석 결과를 보면 논문제목을 통한 데이터의 연계는 제목간의 맵핑 성공률은 약 28.27%(19,370/68,515*100)였다.

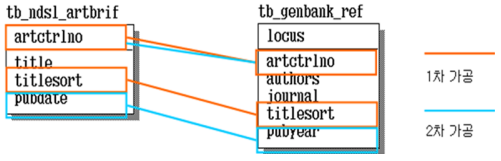

그림 6. 논문 맵핑 데이터 검증

저널 발행연도를 추가로 비교하여 검토한 결과 전체 중복제거 논문건수를 기준으로 약 24.70%의 맵핑 성공률(16,921/68,515*100)을 기록했으며, 정확도는 약 87.36%(16,921/19,370*100)로 나타났다. 위의 작업을 통한 검토 결과 논문 제목을 통한 데이터의 연계는 상당히 높은 정확도(약 87%)를 가지며 데이터의 연계 시에 효과적인 사용이 가능할 것으로 사료된다.

2. NDSL 특허 콘텐츠 연계 방안

Genbank reference의 journal 필드가 patent로 시작되는 데이터를 추출하고 가공하여 locus id, 특허명칭, 출원국가, 특허번호 등의 정보를 분리하고 저장하여 특허 테이블을 생성하였다. [표 3]은 국가별 특허건수 현황이다.

표 3. 국가별 특허건수

국가코드	건수	비율	중복제거	비율
WO(WIPO)	2,084,558	46.28%	24,600	28.24%
US	1,262,903	28.04%	31,529	36.19%
JP	848,400	18.84%	25,169	28.89%
EP	304,433	6.76%	5,823	6.68%
기타	3,814	0.08%	-	
합계	4,504,108	100%	87,121	100%

특허 데이터는 'Patent:' 로 시작되며 국가코드+스페이스+특허번호+스페이스+공개일(등록일)+세미콜론(;)+출원인으로 구성되어 있다(예) Patent: US6,133,00817-OCT-2000;NewEnglandBiolabs,Inc.;Beverly,MA;USA;). 특허 데이터의 국가코드, 특허번호, 공개일(등록일), 공개연도(등록연도)를 분리하여 별도의 컬럼에 저장한 후 NDSL 특허와 연계되도록 하였다. 특허번호의 길이에 따른 데이터 샘플을 10개 이상 분포된 데이터에 대해 KISTI의 NDSL 특허 데이터베이스와 비교한 결과 Genbank 참고문헌 데이터의 가공을 통해 데이터 맵핑이 가능하다는 것을 알 수 있었다.

[표 4]는 Genbank에서 추출한 4,504,108건의 특허 데이터에서 중복을 제거하고 특허번호 자리 수별 현황을 분석하여 특허번호에 따라 가공 가능한 방법을 정리한 내용이다. 한 결과이다. [표 4]에서 보면 자리수의 대부분이 7자리, 8자리, 10자리에 분포되어 있다는 것을 알 수 있다.

특허 데이터 연계를 위해 [표 4]의 방법에 의해서 데이터를 가공하여 [표 5]의 테이블(tb_genbank_mapping)에 로딩하였다. 맵핑 건수는 4,500,217건이며, 전체 건수 대비 약 99.9%의 데이터가 맵핑 가능하였다. 맵핑 테이블의 flag 값에는 P를 입력하면 된다.

표 4. 국가별 특허번호에 따른 가공 방법

구분	자리수	맵핑방법	적용건수	비고
WO(WIPO)	7	현재 상태 링크 가능	803,261	공개번호
	8	첫 자리 숫자 제거	302,879	"
	10	처음 3자리 숫자 제거	978,415	"
	소계		2,084,555	
US	7	현재 상태 링크 가능	1,262,831	등록번호
JP	10	공개일자+공개번호 = JPA공개연월 +' 0' + 특허번호 뒤에서 6자리 사용해서 조회 가능	848,400	NDSL관리번호 검증 검색 시에는 뒤에서 7자리
EP	7	공개번호 = EP-앞에서 7자리	301,650	공개번호
	8	공개번호 = EP-앞에서 7자리	2,781	"
	소계		304,431	
합계			4,500,217	

※ 일본 특허(JP)의 경우 공개번호 만으로 맵핑시킬 경우 1:n의 결과가 나오며 이를 해결하기 위해서 NDSL 관리번호로 가공한 결과를 저장함.

표 5. 맵핑 테이블(tb_genbank_mapping) 구조

컬럼명	데이터타입	크기	Null(Y/N)
locus	varchar	15	N
ndsl_id	varchar	30	N
flag	char	1	N

3. 이종의 콘텐츠 연계 결과

Genbank reference 필드에서 논문의 연계를 위해 데이터베이스를 분석한 결과는 다음과 같다. 중복을 제거한 논문은 68,515건이었고, 중복 제거 논문 중 NDSL 제목과 일치하는 것은 19,370건이었다. 이로써 16,921건이 NDSL 저널 발행연도와 Genbank 저널 정보의 발행연도를 일치한다는 결과를 얻었다. 위에서 기술한 바와 같이 논문제목을 통한 데이터의 연계 맵핑 성공률은 약 28.27%(19,370/68,515*100), 발행년도까지 비교하여 연계한 결과는 24,7%(16,921/68515*100)였다.

Genbank reference 필드에서 특허의 연계를 위해 데이터베이스를 분석한 결과는 논문보다 훨씬 높게 나타났다. 총 4,504,108건의 특허 데이터 중에서 4,500,217건이 맵핑에 성공하여 99.9%의 맵핑되었음을 검증하였다.

IV. 결 론

지금까지 Genbank 데이터베이스를 활용하여 각 필드를 분석하여 그 결과를 산출해 보았으며, 산출된 결과를 기본으로 Genbank와 NDSL 논문, NDSL 특허 연계를 위한 맵핑 테이블을 설계하고 연계 메타 데이터베이스를 구축하였다. Genbank와 NDSL 논문간의 연계를 위해 Pubmed id와 제목, 발행연도를 이용한 데이터 맵핑을 시도하였고, Genbank와 NDSL 특허간의 연계를 위하여 Genbank의 필드를 가공하여 특허번호를 추출하였다.

본 논문에서 분석한 결과를 바탕으로 시스템을 개발하여 서비스한다면 기존 Genbank에서 제공되지 않는 pubmed id 미 보유 논문 중 상당수의 논문이 NDSL 연계로 제공 가능할 것이다. 또한 미국 특허 위주로 서비스되고 있는 Genbank 특허 필드와 KISTI에서 보유하고 있는 유럽 및 일본 특허 정보와의 연계가 가능할 것이다. 이런 콘텐츠간 연계 서비스가 이루어진다면 생명과학 분야 연구자들에게 더욱 유용한 고부가가치 정보를 제공할 수 있을 것으로 기대된다.

참 고 문 헌

[1] 안부영, 한정민, 한건, 이상호, “생명정보 연계검색 인터페이스 설계에 관한 연구”, 제29회 한국정보처리학회 춘계학술발표대회 논문집, 제15권, 제1호, pp.407-409, 2008.

[2] 이상기, 최희윤, 이태석, 한희준, 현미환, 예용희, 김선태, “이종 학술콘텐트 간 연계 및 융합 사례 연구 : KISTI CLICK 중심”, 한국비블리아학회 제19권, 제1호, 2008.

[3] 안부영, 오충식, *생명정보 콘텐츠 업데이트가이드 v. 2.0*, 한국과학기술정보연구원, 2008.

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

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

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

저 자 소 개

안 부 영(Bu Young Ahn) 정회원

- 2003년 8월 : 공주대학교 교육정보대학원(교육정보학 석사)
- 2009년 2월 : 충남대학교 문헌정보학과(문헌정보학 박사)
- 1982년 11월 ~ 현재 : 한국과학기술정보연구원 선임기술원

<관심분야> : 비문헌(사실, 과학) 정보, 메타데이터

이 명 선(Myung Sun Lee) 정회원

- 1996년 2월 : 한남대학교 컴퓨터공학과(공학석사)
- 2005년 2월 : 한남대학교 컴퓨터공학과(공학박사)
- 1983년 3월 ~ 현재 : 한국과학기술정보연구원 책임연구원

<관심분야> : 정보보안, 정보처리, 정보통신

김 지 영(Ji Young Kim) 정회원

- 1997년 2월 : 충남대학교 화학과(이학사)
- 2000년 2월 : 충남대학교 화학과(이학석사)
- 2000년 5월 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 정보서비스, 사실정보, 이용자연구

오 충 식(Chung Shick Oh) 정회원

- 2004년 2월 : 충북대학교 전자계산학과(이학석사)
- 2009년 2월 : 충북대학교 컴퓨터공학과(박사 수료)
- 1986년 3월 ~ 현재 : 한국과학기술정보연구원 선임기술원

<관심분야> : 정보보호, 유비쿼터스, DR