생명정보 콘텐트 업데이트에 관한 연구

안부영, 한정민, 홍순찬, 이상호
한국과학기술정보연구원 사실정보팀
e-mail: {ahnyoung, goal, schong, shlee}@kisti.re.kr

A Study on Update of Bioinformatics Contents

Bu-Young Ahn, Jeong-Min Han, Soon-Chan Hong, Sang-Ho Lee
Factual Information Team,
Korea Institute of Science & Technology Information

요 약

생명과학 기술의 급속한 발달로 인류 복지 증진에 많은 기여를 하였지만 아직도 각종 질병 등으로 많은 사람들이 고통 받고 있으며, 이를 극복하기 위한 연구 및 기술개발은 세계 각처에서 계속되고 있다. 이러한 연구 및 기술개발의 결과로 산출되는 생명정보 데이터의 양은 기하급수적으로 증가하고 있기에 이런 방대한 양의 생명정보 데이터를 분석하고 분석된 데이터에서 인류 복지에 유용한 정보를 얻기 위한 생명정보학(Bioinformatics)이 등장하게 되었다. 이에, 한국과학기술정보연구원(KISTI)은 IT 기반 생명정보 인프라 구축의 중심기관으로 CCBB(Center for Conputationa Biology & Bioinformatics) 웹사이트를 운영하고 있다. CCBB는 전산학적인 기술을 이용한 생명현상 연구를 지원하기 위하여 21종의 생명정보 콘텐트(DB 및 분석도구)를 수집·분석·구축·제공하고 있다. 이 중에서 GenBank, PDB, PIR, Swiss-prot 등의 데이터베이스는 KISTI에서 개발한 KRISTAL 검색엔진을 통하여 국내에서도 빠르고 쉽게 검색 가능하도록 자체 구축하고 있으며, 이와 더불어 BLAST, FASTA, ClustalW 등의 주요 분석 도구 또한 제공하고 있다. 본 논문에서는 CCBB에서 제공중인 21종의 콘텐트 중에서 GenBank, REBASE, GeneCards, InterProScan 등 4종의 대용량 고효율 생명정보 콘텐트의 소개 및 업데이트 방법에 관한 내용을 기술하고자 한다.

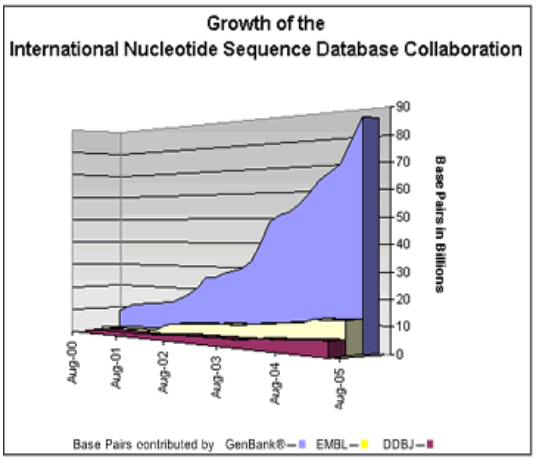
1. 서론

최근 들어 생명과학 연구가 매우 활발하게 진행되고 있으며 이 연구개발의 결과로 나온 생명정보의 양은 기하급수적으로 늘어나고 있는 실정이다. 이런 대용량, 고효율의 생명정보 데이터를 최신의 정보기술을 이용하여 수집·분석 및 데이터베이스로 구축하여 서비스하는 IT기반 생명정보 연구 또한 매우 활발하게 진행되고 있다.

이의 일환으로 미국 NCBI의 GenBank, 영국 EBI의 EMBL(European Molecular Biology Laboratory nucleotide sequence database), 일본의 DDBJ(DNA Data Bank of Japan) 등에서는 각국에서 수집한 유전체 정보를 데이터베이스로 구축하여 실시간 미러로 운영하고 있다. 또한 이러한 유전체, 단백체 정보에 대한 검색과 분석에 대한 다양한 종류의 전산학적 분석도구가 개발되어 배포되고 있다.

매년 기하급수적으로 유전체, 단백체 등 생명정보 관련 데이터가 늘어나고 있다. 그 실례로 (그림 1)은 GenBank, EMBL, DDBJ가 상호 협력하여 구축한 핵산 서열 데이터베이스의 성장현황을 볼 수 있는 그래프이다. 2007년 9월 현재 GenBank 서열 데이터베이스의 구축건수는 76,146,236건(release 161)이다.

한국과학기술정보연구원(KISTI)에서도 기관에서 보유하고 있는 고성능 컴퓨터 및 초고속 네트워크를 토대로 IT 기반 생명정보 인프라 구축의 중심기관으로써 세계 주요 생명정보 데이터베이스 및 분석도구를 수집하여 자체 개발한 KRISTAL 검색엔진을 통하여 국내 연구개발자들이 쉽고 빠르게 이용 가능하도록 전문적인 생명정보 포털 서비스를 제공하고 있다.



(그림 1) 핵산 서열 데이터베이스 구축 현황
(http://www.ncbi.nlm.nih.gov/Genbank)

2. CCBB 콘텐트 현황

KISTI CCBB(http//www.ccbb.re.kr)에서는 GenBank, PDB, Swiss-Prot 등을 비롯한 세계 주요 데이터베이스를 국내에서 이용할 수 있도록 자체적으로 구축함은 물론 BLAST, FASTA, ClustalW 등 다양한 서열 분석 서비스를 제공하여 국내 연구자들의 연구 수행에 필요한 광범위하고도 전문적인 콘텐트를 서비스하고 있다. 특히, 미국 등 여러 선진국과 비교하여 국내 바이오인포매틱스 연구를 위한 인적·물적 자원은 매우 부족한 실정이다. 그래서 KISTI CCBB에서는 고성능 Unix SMP, Linux 클러스터 서버를 기반으로 콘텐트의 지속적인 업데이트 및 유지보수를 통해서 신속하고 정확한 정보를 연구자들에게 제공하고자 노력하고 있다. 더 나아가서 관련 연구자들이 네트워크를 통해 직접 시스템을 활용한 연구를 수행할 수 있도록 기술지원 체제를 마련하여 수행하고 있다. CCBB에서 구축하여 제공하는 콘텐트는 <표 1>과 같다.

<표 1> CCBB 콘텐트 현황(2007/09)

DB 명	구축 건수	비 고
GenBank	76,146,236	release 161
REBASE	4,721	release 708
Ensembl	156,500,000	313Gb(1건=2Kb)
PDB	46,506	2007-09-01
PIR	285,376	-
Swiss-Prot	277,883	release 54.1
CATH	48,391	-
MHCBN	78,000	-
DIP	48,902	-
BIND	80,933	-
GeneCards	48,393	release 2.36
SCOP	103,529	release 1.71
Pfam	9,318	release 22.0
InterProScan	4,949,164	release 16.0
PhiPsi	27,188	-
2Dgel vips	3,052,919	-
담수어류외 20종	27,200	-
합계	**241,733,659**	

3. 콘텐트 업데이트

생명정보 콘텐트(DB 및 분석도구)는 주별, 월별 또는 분기별 등의 다양한 주기로 업데이트 작업이 수행되거나 배포판이 발표된다. 생명정보 데이터베이스의 최신성을 유지하고 사용자들에게 보다 더 정확한 데이터를 제공하기 위해서는 꾸준한 업데이트 작업 및 유지보수 작업을 수행하여야 한다. 지금부터는 GenBank, PDB 등 몇 개의 데이터베이스에 관한 간단한 소개와 함께 업데이트 방법을 기술하도록 하겠다. 워낙 대용량 데이터베이스이다 보니 개인 연구자들이 개인용 컴퓨터로 업데이트를 수행하기에는 불가능할 정도로 시간과 노력이 많이 소요된다.

3.1 GenBank

GenBank는 대표적인 유전자 정보 데이터베이스로써 미국 NCBI(National Center for Biotechnology Information)에서 운영하고 있다. 인간유전체프로젝트(Human Genome Project)의 결과를 포함하여 세계 각지에서 생산된 염기와 단백질 서열정보 및 주석(annotation) 정보를 제공한다. CCBB에서는 KISTI에서 개발한 KRISTAL 검색시스템을 도입하여 데이터베이스를 구축함으로써 국내 생물학 관련 연구자들에게 좀 더 효율적이고 빠른 검색 서비스를 제공하고 있다.

GenBank는 워낙 대용량 데이터베이스인지라 업데이트에 소요되는 시간은 짧게는 3일에서 7일까지 걸린다. 먼저 NCBI GenBank 웹사이트에서 원시 데이터를 FTP를 이용하여 전체 100GB 정도의 원시데이터를 다운로드 받은 후 미리 작성된 KRISTAL 스키마에 맞도록 파일포맷을 변환하여야 한다. 스키마는 xml 형식으로 작성되어 있으므로 상호교환에 어려움이 없다. GenBank 파일형식을 KRISTAL 형식으로 변환하여 KRISTAL 검색엔진을 통하여 검색이 가능하도록 색인작업을 하는 데는 약 60여 시간이 소요된다. 색인작업을 위한 디스크도 300-600GB가 필요하다.

```
[genbank@cc6gene converter]$ more insert_to_schema.xml
<!-- insert this part to db.xml -->
<CreateTable table-name="bct001" with-schema="schema01"/>
<CreateTable table-name="con001" with-schema="schema01"/>
<CreateTable table-name="env001" with-schema="schema01"/>
<CreateTable table-name="est001" with-schema="schema01"/>
<CreateTable table-name="est002" with-schema="schema01"/>
<CreateTable table-name="est003" with-schema="schema01"/>
<CreateTable table-name="est004" with-schema="schema01"/>
<CreateTable table-name="est005" with-schema="schema01"/>
<CreateTable table-name="est006" with-schema="schema01"/>
<CreateTable table-name="est007" with-schema="schema01"/>
<CreateTable table-name="est008" with-schema="schema01"/>
<CreateTable table-name="est009" with-schema="schema01"/>
<CreateTable table-name="est010" with-schema="schema01"/>
<CreateTable table-name="est011" with-schema="schema01"/>
<CreateTable table-name="gss001" with-schema="schema01"/>
<CreateTable table-name="gss002" with-schema="schema01"/>
<CreateTable table-name="gss003" with-schema="schema01"/>
<CreateTable table-name="gss004" with-schema="schema01"/>
<CreateTable table-name="gss005" with-schema="schema01"/>
<CreateTable table-name="htc001" with-schema="schema01"/>
<CreateTable table-name="htg001" with-schema="schema01"/>
<CreateTable table-name="inv001" with-schema="schema01"/>
<CreateTable table-name="mam001" with-schema="schema01"/>
<CreateTable table-name="pat001" with-schema="schema01"/>
<CreateTable table-name="pat002" with-schema="schema01"/>
<CreateTable table-name="phg001" with-schema="schema01"/>
```

(그림 2) GenBank 스키마(xml 형식)

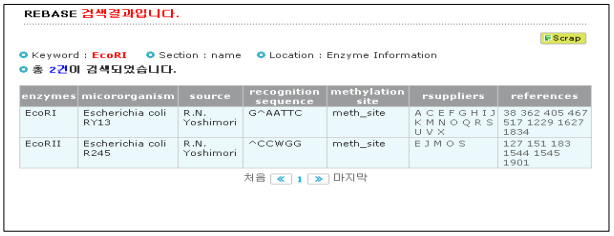
이렇게 작업이 끝난 파일을 KRISTAL에 적재한 후, 프로세스가 제대로 수행되는지 확인한 후 검색 테스트를 거쳐 이용자들이 검색할 수 있도록 서비스 서버를 오픈한다. 검색서비스가 중단되면 안 되기에 서버 2개를 스위칭 방식으로 활용하고 있다.

```
genbank   4806     1  0 Jun07 ?        00:00:00 kristald -D schema/gb.daemon.xml
genbank   4807  4806  0 Jun07 ?        00:00:02 kristald -D schema/gb.daemon.xml
genbank   4808  4806  0 Jun07 ?        00:00:00 kristald -D schema/gb.daemon.xml
genbank   4809  4806  0 Jun07 ?        00:00:00 kristald -D schema/gb.daemon.xml
genbank   4810  4806  0 Jun07 ?        00:00:00 kristald -D schema/gb.daemon.xml
```

(그림 3) GenBank 프로세스 확인 화면

3.2 REBASE

REBASE(The Restriction Enzyme Database)는 제한효소와 그와 관련된 단백질의 정보를 모아 놓은 데이터베이스이다. 제한효소란 DNA에 존재하는 특정한 염기서열을 인식하여 인식한 서열 부위 또는 그 근처에 존재하는 특정부위를 절단하는 단백질이다. 제한효소에는 EcoR Ⅰ, HindⅢ 등이 있는데, 이러한 이름은 처음 분리된 세균의 이름에서 따온 것이다. 제한효소를 이용하여 유전체상에서 제한부위 지도(restriction map)와 같은 DNA 물리적 지도를 만들 수 있으며, 인식 부위와 절단 부위의 특이성이 있는 제한효소는 분자생물학 연구에 유용하게 이용된다.



(그림 4) REBASE 검색결과(효소명: EcoRI)

REBASE에는 제한효소 인식부위와 절단부위, 상업적 활용도, 메틸화 민감도, 결정정보와 서열정보 등이 포함되어 있으며, CCBB에서는 효소정보, 문헌정보, 제한효소를 공급하는 회사 정보별로 주어진 검색조건을 이용하여 제한효소 단백질관련 검색이 가능하다.

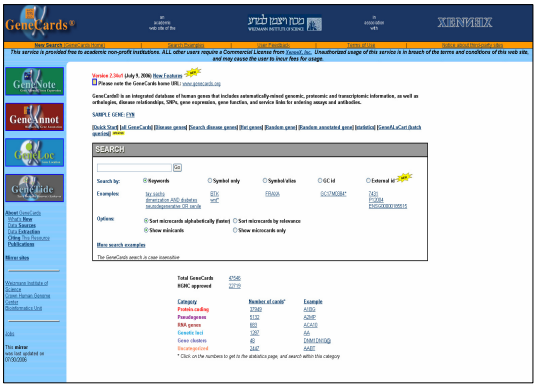
REBASE는 1개월 단위 또는 가변적으로 업데이트 버전을 발표하고 있으며, 2007년 9월 현재 708 버전으로 4,721건의 데이터가 구축되어 서비스되고 있다. 업데이트를 위하여 먼저 NEB(ftp.ftp.neb.com/pub/rebase/*)에서 FTP를 이용하여 데이터를 다운로드 받은 후 필요한 내용(commdata.XXX, parsrefs.XXX)을 선택하여 명령어 라인을 추가한다. 업데이트 프로그램에서 업데이트 날자, 버전, 파일명 등의 변수를 세팅하여 준다. 업데이트 프로그램은 php로 개발하였으며 MySQL 데이터베이스를 활용하고 있다.

<표 2> REBASE enzyme 테이블(MySQL)

필드명	type	비고
name	varchar2($e1)	not null primary key
prototype	char($e2)	
micro_org	varchar2($e3)	
source	varchar2($e4)	
rec_seq	varchar2($e5)	
nocaret	varchar2($e5)	
meth_site	varchar2($e6)	
comm	varchar2($e7)	
ref_nos	varchar2($e8)	

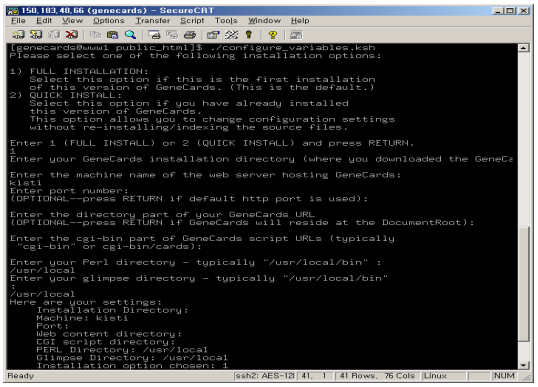
3.3 GeneCards

GeneCards는 이스라엘의 Weizmann연구소에서 개발한 것으로, 인간의 질병에 관련된 유전자 데이터베이스이다. Weizmann연구소와 KISTI는 업무협약을 체결하여 Genecards를 미러사이트로 CCBB에서 서비스하고 있다.



(그림 5) GeneCards 검색화면

GeneCards는 웹 기반의 생물 의학적 정보 검색의 효율성을 높이기 위하여 개발되었으며, 공인된 유전자 이름의 명명법을 기준으로 인간 유전자와 관련된 정보들이 저장되어 있는 GDB(Genome Database), MGD(Mouse Genome Database). OMIM(Online Mendelian Inheritance in Man) 등의 데이터베이스들을 통합하여, 찾고자 하는 유전자 정보를 정리하여 보여준다. 유전자 이름의 명명법은 HUGO(Human Gene Nomenclature Committee)의 기준을 따른다. GeneCards는 상호작용(Interactive) 모드로 환경설정을 마치면 자동 인스톨 가능토록 지원하고 있다.



(그림 6) GeneCards 설치과정

GeneCards는 세계 여러 나라에서 미러사이트로 운영되고 있다. 신규 버전이 나오면 각 미러사이트 담당자들에게 연락이 오고, 연락을 받으면 신속한 업데이트를 수행한다. 업데이트를 위해 먼저 기존 파일을 다른 디렉토리로 이동시킨 후 Weizmann연구소 서버에서 FTP를 이용하여 업데이트할 파일을 다운로드 받는다. 다운로드 받은 후에는

디렉토리 등을 신규 버전에 맞게 수정하고, mirrorActivate.pl, editVars.pl 등의 프로그램이 실행되도록 권한을 변경해 준다. mirrorActivate.pl을 실행하여 환경설정을 수행하고, cards_usr 디렉토리 생성 후 entries를 생성한다. 모든 작업이 완료되면 CCBB 웹사이트에 링크하고 검색을 수행하여 업데이트가 제대로 되었는지 확인한다.

3.4 InterProScan

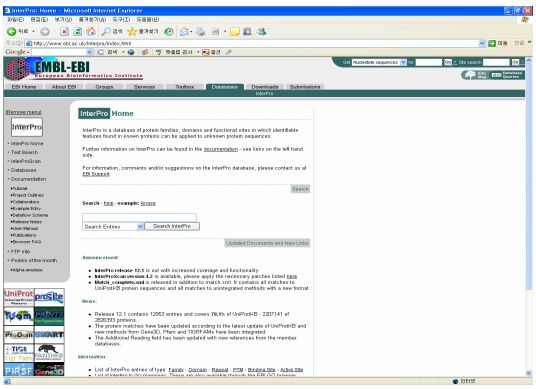
InterProScan은 단백질 도메인과 기능적인 부위(functional sites)에 대한 정보를 모아 놓은 데이터베이스로써 신규 단백질의 기능을 예측하는데 널리 사용되고 있다. 지금까지 알려진 단백질 관련 UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily(PIRSF), SUPERFAMILY 등의 데이터베이스를 통합하였기에 한번의 단백질 서열 검색으로 다양한 결과를 얻을 수 있는 기능을 제공한다. InterProScan에서 지원하는 단백질 관련 데이터베이스의 대략적인 목록은 <표 3>과 같다.

<표 3> InterProScan DB 목록

DB 명	내 용
BlastProDom	▪ ProDom내에 있는 family들을 스캔 ▪ ProDom은 psi-blast를 사용해서 Swiss-prot과 TrEMBL로부터 자동적으로 생성된 단백질 도메인 family들의 포괄적인 set
FPrintScan	▪ PRINTS DB내에 있는 fingerprint들에 대해서 스캔함. ▪ fingerprint들은 단독으로 존재하기보다는 여러 개가 함께 존재하는 것이 우세한 모티프들의 그룹
HMMPIR	▪ 기능적으로 주석이 달인 단백질 서열들인 PIR-PSD로 이루어진 PIR 단백질 서열 DB내에 존재하는 hidden makov model들을 스캔
HMMPfam	▪ PFam 단백질 family DB내에 존재하는 hidden markov model들을 스캔
HMMSmart	▪ Smart 도메인/도메인 family DB내에 존재하는 hidden markov model들을 스캔
HMMTigr	▪ TIGRFAMs 단백질 family DB내에 존재하는 hidden markov model들을 스캔
ProfileScan	▪ PROSITE profile들에 대해서 스캔 ▪ profile들은 weight matrix 기반으로서, 분기된 단백질 family들의 탐지에 민감
ProfileScan	▪ PROSITE 단백질 families&domains DB내에 있는 규칙적인 발현들에 대해서 스캔
SuperFamily	▪ 구조가 알려진 모든 단백질을 나타내는 profile hidden Markov model의 라이브러리

InterProScan을 업데이트 하려면 먼저 EBI 웹사이트에 접속하여 데이터를 다운로드 받는다. 다운로드 한 데이터를 /data/home/iprscan/data/ 디렉토리에 압축을 해제하고 압축해제가 제대로 되었는지 확인한다.

그리고 /data/home/InterProScan/ipscan/data/ 디렉토리의 모든 파일을 Backup_data 폴더로 이동시킨 후 신규 데이터를 서비스 위치로 이동시킨다. 이동시킨 후에 데이터 색인작업을 수행하고 업데이트가 정확하게 되었는지 확인하기 위하여 http://www.ebi.ac.uk/interpro/index.html에서 업데이트 파일 갯수를 확인하여야 한다.



(그림 7) InterProScan 업데이트 확인화면

4. 결론 및 향후 연구방안

생명정보학(Bioinformatics)은 기초생물학, 의학, 응용생물학 분야에 있어서 필수적인 연구수단이고, 생물학, 전산학, 수학, 물리학 등 타 과학영역간의 연계를 기반으로 하는 연구이다. 또한 미래 산업의 주축이 될 생명산업은 인간의 질병 진단과 치료, 신약개발의 핵심적인 기술개발에 주력을 할 것이므로 생명정보 연구개발은 향후 학문과 산업적으로 가장 중요한 분야 중 하나가 될 것이다.

CCBB에서 제공하고 있는 Genbank 등 주요 생명정보 데이터베이스와 BLAST 등 생명정보 분석도구는 세계적으로 가장 많이 사용되고 있는 생명정보 콘텐트이다. 이런 콘텐트를 이용하고자 하는 국내 연구자들에게 최대한 빠른 시간에 최신의 정보와 기술을 제공하고 지원하기 위하여 웹 로봇 등을 개발하여 업데이트를 자동화하려고 추진중이다. 이런 서비스를 제공함으로써 국내 생명과학 연구 효율을 높여 선진국과의 연구개발 격차를 줄여 나가고, 국가 생명과학 연구개발에 기여토록 할 것이다.

참고문헌

[1] 안부영외 “CCBB 서비스 이용자 지침서” 2004. 한국과학기술정보연구원

[2] Evgeni M. Zdobnov, Rodrigo Lopez, Rolf Apweiler, and Thure Etzold. "The EBI SRS server－recent developments" Bioinformatics. 18: 368-373, 2002.

[3] TA Tatusova, I Karsch-Mizrachi, and JA Ostell. "Complete genomes in WWW Entrez: data representation and analysis" Bioinformatics 15: 536-543; 1999.

[4] http://www.ccbb.re.kr

[5] http://www.ncbi.nlm.nih.gov/Genbank