생명정보 콘텐트 구축 매뉴얼

송치평, 이 식, 홍순찬, 이상주

KiSTi www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

생명정보 콘텐트 구축 매뉴얼

KiSTi www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

생명정보 콘텐트 구축 매뉴얼

초판 인쇄: 2005년 12월 12일
초판 발행: 2005년 12월 12일

지은이 | 송치평, 이 식, 홍순찬, 이상주
펴낸이 | 조영화

주소 | 대전시 유성구 어은동 52번지 한국과학기술정보연구원
전화 | (042) 828-5095
팩스 | (042) 828-5179
www.ccbb.re.kr

Published by Center for Computational Biology and Bioinformatics, KISTI
Printed in Republic of Korea

이 책에 대한 의견이나 조언을 주시고자 할 때 그리고 오탈자나 버그 등을 발견했을 경우, 언제든지 저자 중 한 명에게 이메일로 연락주시기 바랍니다.

{chicando,siklee,schong,lsj}@kisti.re.kr

ISBN 89-5884-458-2 93560

- 목 차 -

Ⅰ. 생명정보 DB 구축

1. GenBank

GenBank 데이터베이스는 미국의 NCBI(National Center for Biotechnology Information)에서 운영하는 유전자 정보 데이터베이스이다. GenBank 는 Human Genome Project 결과를 포함하여 세계 각지의 연구실에서 생성된 DNA 염기서열 정보들을 모아 놓은 것으로, 1992 년 10 월 처음으로 서비스를 시작하였다. 현재 GenBank 는 DNA 염기서열 정보와 단백질 정보를 함께 제공하고 있다. 국제적으로 유전자 정보를 저장하는 주요한 데이터베이스로는 GenBank 이외에도 일본의 DDBJ(DNA Data Bank of Japan)와 유럽의 EMBL(European Molecular Biology Laboratory)이 있는데, 이들은 '국제 염기서열 데이터베이스 연합(International Nucleotide Sequence Database Collaboration)'을 통하여 거의 동일한 정보를 공유하고 있으며, 매일 데이터 상호교환을 하여 데이터의 최신성을 유지하고 있다. 2 개월 단위로 업데이트가 진행되고 있으며 2005 년 11 월 현재 150 번째 배포판이 발표된 상태이다.

다음은 GenBank 에서 제공하는 기본적인 파일 형식이다. 각 엔트리와 레코드는 아스키 형태로 자세한 유전자정보를 포함하고 있으며, 특히 레코드는 특수한 표준 형식을 따라 작성되어 있다.

```
❶ LOCUS       XELSRCC2                 115 bp    mRNA    linear   VRT 28-APR-1993
❷ DEFINITION  X.laevis Rous sarcoma virus transforming protein mRNA, 3' end.
❸ ACCESSION   M30860
  VERSION     M30860.1  GI:214812
❹ KEYWORDS    transforming protein.
  SEGMENT     2 of 2
❺ SOURCE      Xenopus laevis (African clawed frog)
    ORGANISM  Xenopus laevis
              Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
              Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae;
              Xenopodinae; Xenopus; Xenopus.
❻ 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
❼ COMMENT        Original source text: X.laevis (female) erythrocyte, cDNA to mRNA.
❽ 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"
❾ ORIGIN         Undetermined number of bp after segment 1.
        1 ctgcaggcgt tcctggagga ctattttaca gctaccgaac cgcagtacca gcctggggac
       61 aacctttagg cttcgctcat aatcaagaga catgtatagg actcttagga aacag
//
```

<GenBank 플랫파일>

GenBank 플랫파일의 첫째 행은 ❶ 로커스(Locus)행이다. 로커스명은 유전자의 위치(locus)를 표현하기 위해 사용된다. 로커스행의 두 번째 항목은 서열의 길이를 나타내며(예: 115bp), 세 번째 항목은 분자의 형태를 표현한다(예: mRNA). 로커스행의 네 번째 항목은 mRNA 가 linear 한 형태를 띠고 있음을 보여준다. 로커스행의 다섯 번째 항목은 GenBank 의 분과코드로서 계통분류학적 의미나 혹은 시퀀싱 기술로 분류하는 목적을 갖는 세 개의 문자이다. GenBank 라이브러리의 분과코드는 다음과 같다. 로커스행의 날짜는 그 기록이 공표되거나 최종적으로 갱신된 날짜를 나타낸다.

<GenBank 라이브러리의 분과 코드>

약어	의미	약어	의미
PRI	영장류 서열	SYN	합성 서열
ROM	설치류 서열	UNA	주석을 달지 않은 서열
MAM	그외 포유동물 서열	EST	EST 서열 (Expressed Sequence Tags)
VRT	그외 척추동물 서열	PAT	특허 서열

INV	무척추 동물 서열	STS	STS 서열(Sequence Tag Sites)
PLN	식물, 균류, 조류 서열	GSS	GSS 서열(Genome Survey Sequence)
BCT	세균 서열	HTG	HTG 서열(High Throughput Genomic sequencing data)
VRL	바이러스 서열	HTC	HTC 서열(High Throughput cDNA sequencing data)
PHG	용균소(bacteriophage) 서열		

❷ DEFINITION 행은 그 자료의 생물학적 특징을 요약하여 나타내는 행이다.

❸ ACCESSION number 는 서열자료의 고유한 식별자로서, 분자 서열에 대응하는 문자와 숫자로 이루어져 있다. ACCESSION 행 하단의 VERSION 행은 등록번호 버전(accession.version)과 유전자정보 검색번호(GenInfo Identifier; GI)를 포함한다. 고유한 염기서열 각각에 대해 이 검색번호가 부여되며, 이것은 데이터베이스의 기록을 조회하는 중요한 요소이다.

❹ KEYWORDS 행은 서열자료의 핵심어를 보여준다. SEGMENT 행은 서열이 여러 부분으로 나눠진 경우, 서열자료가 유전자의 몇 번째 segment 인지를 나타내고 있다.

❺ SOURCE 행에는 서열자료가 얻어진 생물체의 일반명이나 과학명이 기재된다. ORGANISM 에는 서열자료가 유래된 생물의 과학적 학명이 NCBI 분류데이터베이스에서 사용하는 계통 분류에 근거하여 표시된다.

❻ REFERENCE 에서는 인용된 논문의 저자와 논문정보를 보여준다. 또한 MEDLINE 과 PUBMED 의 고유번호도 함께 표시되어 있으며, 이들 데이터베이스와의 직접 링크가 가능하다.

❼ COMMENT 부분은 서열자료에 대한 다양한 주석과 해설을 나타낸다.

❽ FEATURES 부분은 서열자료의 생물학적 특징정보를 보여주는 곳이다. 'Source'에서는 원래의 서열길이, 서열이 유래된 생물체의 과학적인 이름과 분자형태 등의 정보를 보여준다. CDS(coding sequence)에서는 서열 중에서 단백질로 발현되는 부분의 정보를 나타낸다. 위의 GenBank 정보에서 보면, 전체서열길이 115bp 중, 첫 번째에서 69 번째 뉴클레오티드까지의 위치가 CDS 임을 알 수 있다.

❾ ORIGIN 은 서열이 유래한 원래 source 의 전체 서열을 말한다.

1) GenBank DB 다운로드

GenBank DB 파일은 bio-mirror 사이트를 통해서 다운로드 받을 수 있다. 한글 bio-mirror 사이트에 ftp로 접속한 후 다운로드를 수행한다.

DB name	Genbank
ftp address	bio-mirror.kr.apan.net
account	anonymous

$ ftp bio-mirror.kr.apan.net

$ cd pub/biomirror

$ bin

$ get -R genbank

- 다운로드 받을 임의의 디렉토리에서 ftp로 접속한 후 pub/biomirror 위치로 이동한다.
- 다운로드 모드를 binary형태로 바꾸어준다.
- genbank 디렉토리 내의 모든 파일을 다운로드 받는다.(하위 디렉토리 모두 포함)

[작업 수행화면]

```
[web:/data1/genbank/org_data] pwd
/data1/genbank/org_data
[web:/data1/genbank/org_data] ncftp bio-mirror.kr.apan.net
NcFTP 3.0.1 (March 27, 2000) by Mike Gleason (ncftp@ncftp.com).

Copyright (c) 1992-2000 by Mike Gleason.
All rights reserved.

Connecting to 192.249.24.17...
(vsFTPd 2.0.3)
Logging in...
Login successful.
Logged in to bio-mirror.kr.apan.net.
ncftp / > cd pub/biomirror
Directory successfully changed.
ncftp /pub/biomirror > bin
ncftp /pub/biomirror > get -R genbank
tar: No input
genbank/GB_Release_Number.Z:                              8.00 B  130.03 B/s
genbank/README.genbank:                                  13.52 kB  461.40 kB/s
genbank/gbacc.idx.gz:              ETA:   0:40   60.60/475.29 MB   10.40 MB/s
```

2) Bio-KRISTAL 포맷 변환 작업

현재 CCBB에서는 GenBank DB를 한국과학기술정보연구원에서 자체 개발한 Bio-KRISTAL 검색엔진을 이용하여 보다 빠른 검색 속도와 정확도를 선보이고 있다. 이에 따라 다운로드 받은 GenBank DB를 Bio-KRISTAL 검색엔진을 이용하여 서비스하기 위해서는 사전 변환 작업이 필요하다.

▶ 실행중인 Bio-KRISTAL Daemon을 중지 시킨다.

$ cd ~genbank

$ KRISTALd_stop schema/gb.daemon.xml

▶ 다운로드 받은 파일중에서 est 파일들은 EST 디렉토리로 이동시킨다.

$ cd ~genbank/org_data

$ mkdir EST

$ mv gbest* EST

▶ Bio-KRISTAL 포맷에 맞도록 DB convert 스크립트를 실행한다. 24시간 정도 소요.

$ cd ~genbank/converter

$ time sh run.converter.sh

[작업 수행화면]

```
[genbank@genbank genbank]$ cd converter
[genbank@genbank converter]$ sh run.converter.sh
+-----------------------------------------------+
| Converter for GenBank format to KST format    |
+-----------------------------------------------+
  Source Dir. : /home/genbank/org_data
  Data Direc. : /home/genbank/data
  # of Source files : 377

  1 [/home/genbank/org_data/gbbct1.seq.gz] ..... 2497
```

▶ convert 작업을 수행하고 나면 insert_to_schema.xml 파일이 생성된다. 이파일 내용을 /schema 디렉토리에 존재하는 Genbank.db.xml과 Genbank.load.xml 파일에 추가시켜 준다.

▶ 변환된 Bio-KRISTAL 포맷파일을 load 스크립트를 실행하여 적재작업을 수행한다. DB 사이즈에 따라 다르지만, 보통 30시간 이상 시간이 소요된다.

$ cd ~genbank

$ time sh _load &

3) Bio-KRISTAL Daemon 실행 및 에러대처 방법

▶ Bio-KRISTAL Daemon 실행 시키기.

$ cd ~genbank

$ KRISTALd -D schema/gb.daemon.xml

▶ Daemon 실행시 에러대처: Bio-KRISTAL Daemon이 비정상 종료되었거나 종료되지 않은 상태에서 load작업을 수행한 후에 Daemon을 실행시키면 다음과 같은 에러화면이 나타날 수 있다.

```
에러예제:
:> ----------------------------------------------------
PRUNING_SIZE      : 0
----------------------------------------------------
          1 processors are running...
          DB Management processor or Defunct processor..
          Kill all processor and restart(y/n) ?
==> y   /* 'y'을 입력한다. */
```

init_db_env을 실행시켜서 DB관련 환경변수파일들을 clear 시킨후 데몬을 다시 실행한다.

$ ./init_db_env /home/genbank/gb_volumes

2. dbSNP

단일염기다형성(Single Nucleotide Polymorphism: SNP)이란 한 개의 염기서열에서 다형성(ploymorphism)이 나타나는 것을 말하며, 평균적으로 인구 집단의 1% 이상의 빈도로 일어나는 유전적인 변이를 말한다. SNP는 인간 개개인의 표현형의 차이, 특정질환에 대한 감수성, 환경적인 영향에 대한 개체반응의 유전적 특징을 규정짓는 것으로 생각되고 있다. SNP 연구의 중요성은 질병과 관련된 유전자 또는 유전지표의 발견, 유전적인 마커(marker)로서의 가능성, 개인의 유전자형에 따라 특정약물에 대한 반응이나 질환 유발에 대한 민감도가 다르다는 사실을 고려한 맞춤의약 관련연구에 활용될 수 있다는 것이다.

1) dbSNP 원문파일 다운로드 하기

dbSNP는 mysql DB에 적재하여 서비스를 제공한다. 최신 snp DB파일은 ftp로 다운로드 받는 방법과 rsync를 통해서 동기화 할 수 있는 2가지 방법을 제공한다. 이문서에서는 ftp로 다운받는 방법을 이용하여 작업을 수행한다.

DB name	dbSNP
ftp address	ftp.ncbi.nih.gov/snp
account	anonymous

$ cd dbSNP_orgdata

$ ftp ftp.ncbi.nih.gov/snp

$ cd mssql/data

$ bin

$ get -R * .

▶ 다운로드 받은 파일 압축을 해제한다.

$ gunzip *.gz

2) 업데이트 수행하기

dbSNP는 perl로 작성된 update_snp.pl을 실행하여 업데이트 작업을 수행한다.

```
#!/usr/bin/perl

$fileNameList = $ARGV[0];

open (LIST,"$fileNameList") or die "cannot open $fileNameList \n";
while ($list = <LIST>){

    chop($list);
    $fileName = $list;
    ($table,$t)= split /\./,$fileName,2;
    system("/usr/local/mysql/bin/mysql -u root -pdptm5173 -e \"delete from $table ;\" dbSNP ");
    system ("/usr/local/mysql/bin/mysqlimport -u root -pdptm5173 dbSNP  $fileName ");

}
close(LIST);
```

$ cd bin

$ update_snp.pl list > log.txt

3) 업데이트 결과 체크 및 버전정보 확인

정상적으로 업데이트 작업이 수행되었는지 체크를 위해 update_check.pl을 실행한다.

$ update_check.pl list > check_log.txt

업데이트 건수 및 버전정보를 확인하기 위해서는 다음 사이트를 참고한다.

http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi

3. REBASE

REBASE(The Restriction Enzyme Database)는 제한효소 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 DNA의 물리적 지도를 만들 수 있으며, 인식 부위와 절단 부위의 특이성이 있는 제한효소는 분자생물학 연구에 유용하게 이용되고 있다.

REBASE는 제한효소와 그와 관련된 단백질의 정보를 모아놓은 것으로 저널에 게재된 것과 게재되지 않은 자료, 제한효소 인식 부위와 절단 부위, 상업적 활용도, 메틸화 민감도(methylation sensitivity), 결정정보(crystal data)와 서열정보를 포함한다.

DNA methyltransferase, homing endonuclease, nicking enzyme, 특정 단위체(subunit)와

조절 단백질 등의 정보도 제공하고 있다. 최근에는 유전체의 서열분석을 통해 추정 DNA methyltransferases와 제한 효소 정보도 보여준다.

1) REBASE 다운로드 및 파일 수정하기

ftp에 접속한 후 /pub/rebase 디렉토리에서 commdata.XXX, parsrefs.XXX 파일을 다운로드 받는다.

DB name	REBASE
ftp address	ftp.neb.com
account	anonymous

$ ftp ftp.neb.com

$ cd pub/rebase

$ bin

$ mget commdata.XXX parsrefs.XXX

▶ commdata.XXX 파일을 편집기로 열어 제일 마지막 부분에 '------'라인을 추가한다. 추가한 라인 다음에는 공백 및 라인이 존재해서는 안된다.

[작업 수행화면]

```
Cambridge Bioscience
24-25 Signet Court, Newmarket Road, Cambridge CB5 8LA
U.K.
Tel:    44 1223 316855
Fax:    44 1223 360732

Clontech
1020 East Meadow Circle, Palo Alto, CA  94303-4280
USA
Tel:    415 424 8222
Fax:    415 424 1088
------------------------------------------------------------------------
~
```

2) 업데이트 수행하기

REBASE 업데이트 프로그램은 CCBB에서 자체 제작한 php 웹어플리케이션을 이용하여 업데이트 작업을 수행한다. 아래 그림에서 보는 바와 같이 table create, drop, insert, insert test, select 기능을 포함하고 있다. 순서대로 기존 테이블을 drop하고 다시 테이블을 생성한 후 insert하는 단계로 업데이트 작업을 수행한다.

▶ 업데이트를 위해 환경변수가 셋팅된 globals.php를 열어 버전에 맞게 내용을 수정한다.

$ update = "Nov 14, 2005"

$ version = “REBASE version 511”

$ parsrefs = “parsrefs.511”

$ refs = “parsrefs.511”

$ commdata = “commdata.511”

[REBASE 업데이트 웹어플리케이션]

주소: ~rebasetest/tables.php

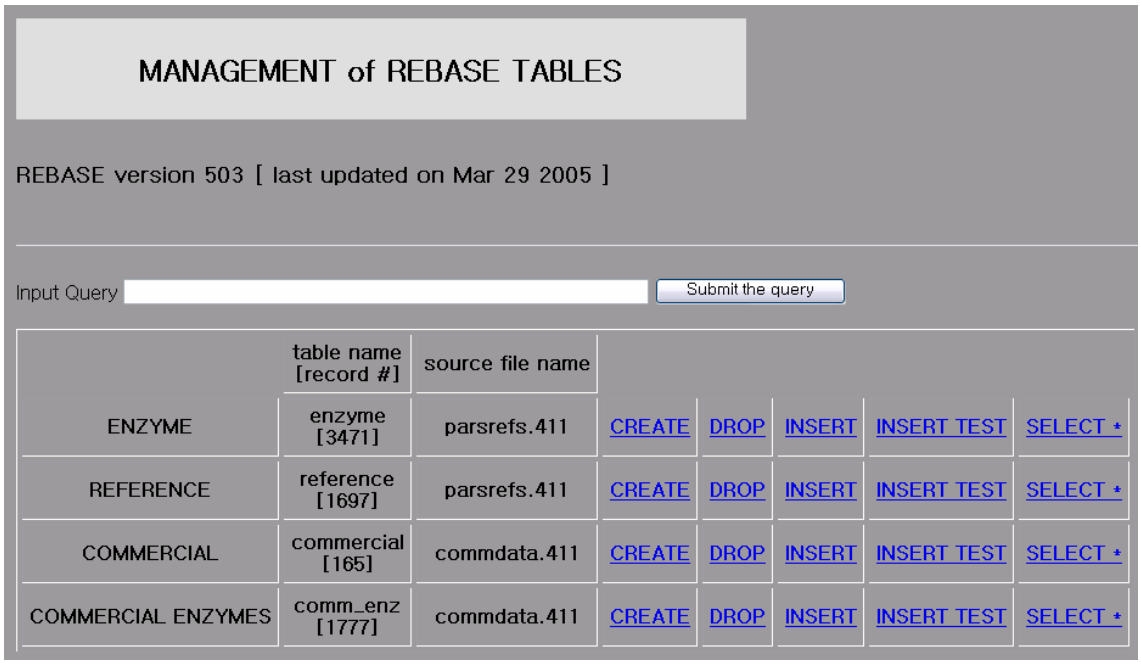


	table name [record #]	source file name					
ENZYME	enzyme [3471]	parsrefs.411	CREATE	DROP	INSERT	INSERT TEST	SELECT *
REFERENCE	reference [1697]	parsrefs.411	CREATE	DROP	INSERT	INSERT TEST	SELECT *
COMMERCIAL	commercial [165]	commdata.411	CREATE	DROP	INSERT	INSERT TEST	SELECT *
COMMERCIAL ENZYMES	comm_enz [1777]	commdata.411	CREATE	DROP	INSERT	INSERT TEST	SELECT *

작업순서: DROP -> CREATE -> INSERT TEST -> INSERT ->SELECT(확인과정)

▶ 업데이트 관련 참고사항

- 혹 추후에 REBASE의 자료가 계속 늘어날 경우, 줄단위로 읽을시 문제가 발생할 소지가 있음.
- 이러한 문제가 발생하는지는 Insert Test를 실행하여 정상적으로 종료하지 확인하여 문제여부를 확인한다.
- 또한 새로운 자료들중 현재 구성된 테이블의 각 필드 데이터형 크기와 일치하지 않아 문제가 발생할 수 있는데, 이는 실행시 오라클의 에러를 보고, globals.php안의 변수를 수정하여 해당 테이블의 데이터정의부분을 수정하면 된다.

4. PDB

PDB(Protein Data Bank)는 생물학적 거대분자의 3차원 구조를 저장해 놓은 국제적인 공공

데이터베이스이다. PDB 자료는 X-ray 회절법(crystallography)과 NMR 실험으로부터 나온 실험 데이터이다. 이 자료들을 기반으로 하여 단백질의 정보와 삼차원 구조 영상 등을 제공한다. CCBB의 PDB 데이터베이스는 기존에 구축되어 있는 RDBMS(Relational Database Management System)에서 벗어나, KISTI의 KRISTAL-2002 정보검색시스템을 기반으로 하여 생물정보자료를 다룰 수 있도록 개발된 Bio-KRISTAL 시스템을 이용하여 구축되었다. PDB 데이터는 1주일마다 갱신되며, 2005년 11월 14일을 기준으로 33,454건의 데이터가 구축되어 있다.

1) PDB DB파일 다운로드 하기

PDB 최신 DB파일을 다운로드 받을 수 있는 방법은 ftp, rsync을 이용할 수 있다. rsync을 이용한 다운로드 방법은 다음과 같다.

```
$ /usr/local/bin/rsync -rlpt -v -z -delete -port=33444
rsync.rcsb.org::ftp_data/data/structures/divided/pdb
pdb_home/data > log.txt 2>/dev/null &
```

이 문서에서는 ftp로 다운받는 방법을 이용하여 PDB 업데이트를 수행하도록 한다.

DB name	PDB
ftp address	ftp.rcsb.org
account	anonymous

```
$ ftp ftp.rcsb.org
$ cd pub/pdb/data/structures/divided/pdb
$ bin
$ get -R *
```

2) 다운로드 받은 원문파일 압축풀기

여러 개의 파일을 한번에 압축을 풀 수 있도록 스크립트를 작성해서 수행한다.

[자동 압축풀기 스크립트 예]

```
gunzip a[0-z]*/*.Z
gunzip b[0-z]*/*.Z
   -- 중략 --
gunzip z[0-z]*/*.Z
gunzip 1[0-z]*/*.Z
gunzip 2[0-z]*/*.Z
gunzip 3[0-z]*/*.Z
gunzip 4[0-z]*/*.Z
gunzip 5[0-z]*/*.Z
gunzip 6[0-z]*/*.Z
gunzip 7[0-z]*/*.Z
gunzip 8[0-z]*/*.Z
gunzip 9[0-z]*/*.Z
gunzip 0[0-z]*/*.Z
```

3) Bio-KRISTAL 포맷으로 변환작업하기

KRISTAL Daemon을 이용한 검색 서비스를 제공하기 위해서는 Bio-KRISTAL 포맷으로 사전 변환작업이 필요하다. C언어로 개발된 포맷 변환 프로그램을 실행시킨다.

$ pdb /source_folder pdb

USEAGE : PDB.EXE SOURCE_FOLDER OUTPUT_FILE

변환작업 결과로 *.kst 파일이 생성되며 이파일들은 /pdb_home/pdb/data 디렉토리로 옮겨준다. 기존의 파일들이 존재하면 새로운 파일로 덮어씌우면 된다.

4) Bio-KRISTAL Daemon 중지하기

변환된 KRISTAL 포맷파일을 Load하기 위해서는 실행중인 KRISTAL Daemon을 중지시켜야 한다.

$ pdb_home/K2000/bin/KRISTALmgr stop /pdb_home/pdb/k2000config/pdb.conf

여기에서 pdb_home에는 실제적인 pdb home 디렉토리를 적어 주면 된다.

5) DB schema 생성 및 DB 업데이트 하기

pdb_home 디렉토리 에서 DB schema을 생성하기 위한 다음 명령어를 실행한다. 실행결과로 /pdb_home/volumes 디렉토리 안에 PDB.CAT, PDB.DICT 파일이 생성된다.

$ Loader -cbi -f ./schema/pdb.schema

성공적으로 수행하고 나면 실제적으로 업데이트 하는 명령어를 다음과 같이 실행한다.
약 20시간 정도의 시간이 소요된다.
$ Loader -abi -f ./schema/pdb_1.schema

6) Bio-KRISTAL Start하기
$ pdb_home/K2000/bin/KRISTALmgr restart /pdb_home/pdb/k2000config/pdb.conf

7) 에러대처 방법
'Loader -abi -f ./schema/pdb_1.schema' 실행시 발생하는 CAT, DICT 파일 존재하지 않을 때 발생하는 에러로써 DB에 대한 카달로그 정보 관련한 파일이 존재하지 않을 때 에러가 발생한다.
[에러 메시지]

```
Starting KRISTAL-2000 Loader...
Invalid DBG Dir:/pdb_home/pdb/volumes
Invalid DBGroup Name:PDB
invalid volume id
*** Error in reading schema info from CATALOG DB
Shutting down KRISTAL-2000 Loader
Finished!
```

'Loader -cbi -f ./schema/pdb.schema' 실행으로 volumes 디렉토리내에 PDB.CAT, PDB.DICT 생성해 주고 다시 Bio-KRISTAL을 start해주면 에러를 해결할 수 있다.

5. PIR

PIR(The Protein Information Resource)은 Georgetown University Medical Center (GUMC)에서 운영하고 있으며 단백질 데이터베이스와 단백질 분석 도구를 제공한다. 단백질 서열에 기능 주석을 붙인 단백질 데이터베이스로 단백질 서열 검색, 도메인 분석 등을 제공하며, 텍스트 검색과 Unique Identifiers, Accessions, Cross References 등을 이용한 검색이 가능하다. CCBB에서 제공하는 PIR 데이터베이스는 기존에 구축되어 있는 RDBMS (Relational Database Management System) 에서 벗어나, KISTI의 KRISTAL-2002 정보검색시스템을 기반으로 하여 생물정보 데이터를 다룰 수 있도록 개발된 Bio-KRISTAL 시스템을 이용하여 구축되었다. PIR 데이터베이스는 PIR-PSD(PIR-International Protein Sequence Database) Release 80 버전을 마지막으로 2004년 12월 31일을 기준으로 283,009건의 데이터가 구축되어 있다.

1) PIR DB파일 다운로드 하기

PIR 최신 DB파일을 다운로드 받을 수 있는 방법은 ftp을 이용한 다운로드 방법을 사용하고 있다.

DB name	PIR
ftp address	pir.georgetown.edu
account	anonymous

$ ftp pir.georgetown.edu

$ cd pir_databases/psd/codata

$ bin

$ get -R *

2) Bio-KRISTAL 포맷으로 변환작업하기

KRISTAL Daemon을 이용한 검색 서비스를 제공하기 위해서는 Bio-KRISTAL 포맷으로 사전 변환작업이 필요하다. C언어로 개발된 포맷 변환 프로그램을 실행시킨다.

$ pir /source_folder pir

USEAGE : PDB.EXE SOURCE_FOLDER OUTPUT_FILE

변환작업 결과로 *.kst 파일이 생성되며 이파일들은 /pir_home/pir/data 디렉토리로 옮겨준다. 기존의 파일들이 존재하면 새로운 파일로 덮어씌우면 된다.

3) Bio-KRISTAL Daemon 중지하기

변환된 KRISTAL 포맷파일을 Load하기 위해서는 실행중인 KRISTAL Daemon을 중지시켜야 한다.

$ pir_home/K2000/bin/KRISTALmgr stop /pir_home/pir/k2000config/pir.conf

여기에서 pir_home에는 실제적인 pir home 디렉토리를 적어 주면 된다.

4) DB schema 생성 및 DB 업데이트 하기

pir_home 디렉토리 에서 DB schema을 생성하기 위한 다음 명령어를 실행한다. 실행결과로 /pir_home/pir/volumes 디렉토리 안에 PIR.CAT, PIR.DICT 파일이 생성된다.

$ Loader -cbi -f ./schema/pir.schema

성공적으로 수행하고 나면 실제적으로 업데이트 하는 명령어를 다음과 같이 실행한다.

약 10시간 정도의 시간이 소요된다.

$ Loader -abi -f ./schema/pir1.schema

5) Bio-KRISTAL Start하기

$ pir_home/K2000/bin/KRISTALmgr restart /pir_home/pir/k2000config/pir.conf

6) 에러대처 방법

'Loader -abi -f ./schema/pir_1.schema' 실행시 발생하는 CAT, DICT 파일 존재하지 않을 때 발생하는 에러로써 DB에 대한 카달로그 정보 관련한 파일이 존재하지 않을 때 에러가 발생한다.

[에러 메시지]

```
Starting KRISTAL-2000 Loader...
Invalid DBG Dir:/pir_home/pir/volumes
Invalid DBGroup Name:PIR
invalid volume id
*** Error in reading schema info from CATALOG DB
Shutting down KRISTAL-2000 Loader
```

' Loader -cbi -f ./schema/pir.schema' 실행으로 volumes 디렉토리내에 PIR.CAT, PIR.DICT 생성해 주고 다시 Bio-KRISTAL을 start해주면 에러를 해결할 수 있다.

6. SWISS-PROT

SWISS-PROT은 단백질 서열 데이터베이스이다. 이곳에서는 단백질의 기능, 도메인 구조나 변이 등과 같은 세부 정보를 검색할 수 있다. CCBB SWISS-PROT 데이터베이스는 기존에 구축되어 있는 RDBMS (Relational Database Management System) 에서 벗어나 KISTI의 KRISTAL-2002 정보검색시스템을 기반으로 하여 생물정보자료를 다룰 수 있도록 개발된 Bio-KRISTAL 시스템을 이용하여 구축되었다. 현재 SWISS-PROT 데이터베이스는 Release 48.4 버전을 토대로 2005년 11월 14일 현재 197,228건의 데이터가 구축되어 있다.

1) SWISS-PROT DB파일 다운로드하기

SWISS-PROT 최신 DB파일을 얻기 위해서 ftp을 이용한 다운로드 방법을 사용하고 있다.

DB name	SWISSPROT
ftp address	ftp.ebi.ac.uk
account	anonymous

$ ftp ftp.ebi.ac.uk

$ cd pub/databases/swissprot/release_compressed/
$ bin
$ get uniprot_sprot.dat.gz

2) 다운로드 받은 원문파일 압축풀기
여러 개의 파일을 한번에 압축을 풀 수 있도록 스크립트를 작성해서 수행한다.
$ gunzip uniprot_sprot.dat.gz

3) Bio-KRISTAL 포맷으로 변환작업하기
KRISTAL Daemon을 이용한 검색 서비스를 제공하기 위해서는 Bio-KRISTAL 포맷으로 사전 변환작업이 필요하다. C언어로 개발된 포맷 변환 프로그램을 실행시킨다.
$ swiss /source_folder swiss
USEAGE : PDB.EXE SOURCE_FOLDER OUTPUT_FILE
변환작업 결과로 *.kst 파일이 생성되며 이파일들은 /swiss_home/swissprot/data 디렉토리로 옮겨준다. 기존의 파일들이 존재하면 새로운 파일로 덮어씌우면 된다.

4) Bio-KRISTAL Daemon 중지하기
변환된 KRISTAL 포맷파일을 Load하기 위해서는 실행중인 KRISTAL Daemon을 중지시켜야 한다.
$ swiss_home/K2000/bin/KRISTALmgr stop
/swiss_home/swissprot/k2000config/swiss.conf
여기에서 swiss_home에는 실제적인 swiss home 디렉토리를 적어 주면 된다.

5) DB schema 생성 및 DB 업데이트 하기
swiss_home 디렉토리 에서 DB schema을 생성하기 위한 다음 명령어를 실행한다. 실행결과로 /swiss_home/swissprot/volumes 디렉토리 안에 SWISSPROT.CAT, SWISSPROT.DICT 파일이 생성된다.
$ Loader -cbi -f ./schema/swiss.schema

성공적으로 수행하고 나면 실제적으로 업데이트 하는 명령어를 다음과 같이 실행한다.
약 10시간 정도의 시간이 소요된다.
$ Loader -abi -f ./schema/swiss_1.schema

6) Bio-KRISTAL Start하기
$ swiss_home/K2000/bin/KRISTALmgr restart /swiss_home/swissprot/k2000config/swissprot.conf

7) 에러대처 방법

'Loader -abi -f ./schema/swissprot_1.schema' 실행시 발생하는 CAT, DICT 파일 존재하지 않을 때 발생하는 에러로써 DB에 대한 카달로그 정보 관련한 파일이 존재하지 않을 때 발생하는 에러.

[에러 메시지]

```
Starting KRISTAL-2000 Loader...
Invalid DBG Dir:/swissprot_home/swissprot/volumes
Invalid DBGroup Name:SWISSPROT
invalid volume id
*** Error in reading schema info from CATALOG DB
Shutting down KRISTAL-2000 Loader
Finished!
```

' Loader -cbi -f ./schema/swissprot.schema' 실행으로 volumes 디렉토리내에 SWISSPROT.CAT, SWISSPROT.DICT 생성해 주고 다시 Bio-KRISTAL을 start해주면 에러를 해결할 수 있다.

7. InterProScan

InterProScan은 단백질 도메인과 기능적인 부위들(functional sites)에 대한 정보를 모아 놓은 데이터베이스로서 신규 단백질의 기능을 예측하는데 널리 사용되고 있다. 지금까지 알려진 단백질 관련 데이터베이스인 UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily(PIRSF), SUPERFAMILY 등의 데이터베이스를 모두 통합해 놓았기 때문에 한번의 단백질 서열 검색으로 다양한 결과를 얻을 수 있는 편리한 기능을 제공한다.

1) InterProScan DB 파일 다운로드 받기

ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/DATA/README 파일에서 현재 발표된 버전을 확인 한 후 업데이트가 필요하면 다음 주소로 접속하여 필요한 파일들을 다운로드 받는다.

DB name	InterProScan
ftp address	ftp.ebi.ac.uk
account	anonymous

$ ftp ftp.ebi.ac.uk
$ cd pub/databases/interpro/iprscan/DATA
$ bin
$ get iprscan_DATA_11.0.tar.gz

2) 다운로드 받은 파일 압축 풀기
$ gunzip -c iprscan_DATA_11.0.tar.gz | tar xvf -

3) 인덱싱 및 적재 작업 수행 하기
ebi에서 제공하는 perl로 개발된 InterProScan 인덱싱과 적재 프로그램을 수행한다. 초기 인덱싱 작업에는 많은 시간이 소요되지만 버전 업그레이드에 따른 작업이므로 2시간 정도의 작업시간이 소요된다.
$ cd ~InterProScan/iprscan/bin
$ index_data.pl

Ⅱ. Mirror Site 구축

1. GeneCards

GeneCards는 이스라엘의 Weizmann 연구소에서 개발한 인간의 질병에 관련된 유전자 데이터베이스 서비스이다. CCBB에서는 고성능 컴퓨팅 시스템 환경에 GeneCards 미러 서비스를 구축하여 연구자들이 좀더 빠른 검색과 분석결과를 얻을 수 있도록 서비스를 제공하고 있으며 2005년 11월 현재 GeneCards 2.33을 서비스하고 있다.

1) Genecards 소스파일 다운로드 받기
Weizmann 연구소에서는 Mirror Site 사전 협약 과정을 거친 후 소스파일을 다운로드 받을 수 있도록 계정을 발급해 준다. 그리고 업데이트가 수행될 때 마다 새로운 패스워드를 발급해서 사용할 수 있도록 지원하고 있다.

DB name	GeneCards
ftp address	miriam.weizmann.ac.il
account	다운로드시 마다 변경됨

$ ftp miriam.weizmann.ac.il
$ bin
$ get source_file

2) 소스 파일 수정하기
다운로드 받은 소스파일 중에 환경변수 내용을 설치 할 시스템에 맞게 수정해 주는 작업이 필요하다.

▶ Mirror_activate 파일 첫번째 라인에 있는 perl 경로를 시스템에 맞게 수정한다.
$>vi mirror_activate
#!/usr/local/perl ---> #!/usr/bin/perl/ -w

▶ configure_variable.ksh, mirror_activate을 실행 할 수 있도록 권한을 변경해 준다.
$ chmod a+x configure_variables.ksh, mirror_activate

3) configure_variable.ksh 실행하여 환경변수 셋팅하기
configure_variable.ksh을 실행하면 인터프리터 형식으로 환경변수 값을 셋팅 할 수 있다.

실행결과로 variables라는 파일이 생성되며 셋팅해야 하는 값들은 다음과 같다.

```
#########################################################################
# VARIABLES THAT YOU MUST EDIT
#########################################################################

# The full pathname of this GeneCards mirror
$basedir_genecards="/data1/genecards/public_html";

# The URL of this GeneCards mirror
$baseURL_genecards_html="http://genecards.ccbb.re.kr/cards";

# The URL of the cgi-bin directory of this GeneCards mirror
# A typical URL for a mirror site would end with '/cgi-bin/cards'.
# 'cards-bin' is an HTTP server ScriptAlias on the master site
$baseURL_genecards_cgi="http://genecards.ccbb.re.kr/cgi-bin";

# The location of Glimpse
$glimpsedir="/usr/local/bin" ;

# The location of Perl
$perldir= "/usr/bin";

# The location of your grep executable
$grepPATH= "/bin";
```

4) cards_usr 폴더 생성후 entries 생성하기

mirror_activate 소스에 따라 가변적인 특성을 가지고 있다.

$ mkdir cards_usr

$ mv gcXML2.31.tar.Z ./cards_usr/.

$ zcat gcXML2.31.tar.Z | tar xof -

$ mv xml_entries entries

5) 마지막으로 mirror_activate을 수행하여 DB 인덱싱 작업과 적재작업을 실행한다.

2. Pfam

Pfam 서버는 다음과 같은 디렉토리 구조를 가지고 있으며 각각의 기능 및 내용의 이해를 돕기 위해 다음과 같이 정리한다.

1) Pfam 서버 구조 이해하기

▶ Pfam 디렉토리, Hmmer 디렉토리, bin 디렉토리

pfamdist: Pfam site에서는 제공하는 릴리즈 파일을 다운 받아 보관하는 디렉토리. 초기에는

empty 상태.
hmmerdist: http://hmmer.wustl.edu에서 제공하는 HMMER code을 다운 받아 보관하는 디렉토리. 초기에는 empty 상태.
bin: HMMER 관련 프로그램이 존재하는 디렉토리(hmmalign, hmmcalibrate, hmmemit, hmmindex, hmmsearch, hmmbuild, hmmconvert, hmmfetch, hmmpfam), 초기에는 empty 상태.

▶ Pfam 서버 설치 및 설정과정에서 생성되는 디렉토리
Desc: 데이터 항목별 Description 파일이 존재하는 디렉토리.
Full: Full alignments, Pfam site에서 다운받은 릴리즈 파일(pfamdist 디렉토리내)을 변환하여 파일을 생성.
Seed: Seed alignments, Pfam site에서 다운받은 릴리즈 파일(pfamdist 디렉토리내)을 변환하여 파일을 생성.
Otherdata: Pfam_fs, Pfam_ls, swisspfam, swissother, domain.pnh 파일이 존재하는 디렉토리, Pfam site에서 다운받은 릴리즈 파일(pfamdist 디렉토리내)을 변환하여 파일을 생성.

▶ 웹서버를 위해 설정과정에서 생성하는 디렉토리
logs: Pfam 서버 로그를 기록하기 위한 디렉토리, Pfam ROOT에 생성.
tmp: 데이터 검색과정에서 발생하는 임시파일 기록을 위한 디렉토리, Pfam ROOT에 생성.
Docs/images/tmp: domain 구조 GIF을 기록하기 위한 임시 디렉토리.

2) Pfam Install 사전작업 하기
▶ Pfam 계정 및 디렉토리 생성
Pfam 서버를 설치하기 위한 계정을 생성하고 디렉토리를 설정한다.

3) Pfam 소스파일 다운로드 하기
ftp.genetics.wustl.edu(anonymous)로 접속하여 Pfam 서버 패키지를 다운 받은 후 다음과 같이 Pfam 계정 디렉토리내에 압축을 해제하여 설치한다.

DB name	Pfam
ftp address	ftp.genetics.wustl.edu
account	anonymous

$ ftp ftp.genetics.wustl.edu
$ cd pub/Pfam
$ bin
$ get pfamserver-17.0.tar.gz

$ bye
$ gunzip pfamserver-17.0.tar.gz
$ tar xf pfamserver-17.0.tar

정상적으로 설치가 이루어지고 나면 /pfam_home/pfamserver-17.0 디렉토리가 생성된다.

4) 서버환경 설정하기
/pfam_home/pfamserver-17.0/conf/pfamconf.pl을 수정하여 서버환경에 맞게 설정한다.

▶ 아파치 설정파일 httpd.conf에 다음과 같은 내용을 추가해 준다.
make a config section for the DocumentRoot:
<Directory /pfam_home/pfamserver-17.0/docs>

make a config section for the CGI directory
<Directory /pfam_home/pfamserver-17.0/cgi-bin

▶ 기타 설정내용

```
pfamconf.pl 예제:
$RELEASE   = "17.0";                          # Used in lots of places
$RELDATE   = "July 2005";                     # Shows up on homepage
$LOCATION = "KISTI";                          # Used in navigation bar
$SERVER    = "pfam.ccbb.re.kr";               # Some URLs must be constructed to
$CONTACT   = "ccbb@kisti.re.kr";              # Where feedback is directed to
$IMAGE     = "ccbb_logo.gif";                 # Pretty picture in the docs/images/
$PFAMROOT = "/pfam_home/pfamserver-17.0";     # Pfam "home directory"
```

5) HMMER, Pfam 릴리즈 파일 다운로드 하기
▶ HMMER 다운로드 절차
$ cd ~pfam_home/pfamserver-17.0/hmmerdist
$ ftp ftp.genetics.wustl.edu(anonymous)
$ cd pub/eddy/hmmer/CURRENT
$ bin
$ get hmmer-2.3.2.bin.intel-linux.tar.gz
$ bye
$ gunzip hmmer-2.3.2.bin.intel-linux.tar.gz

$ tar xf hmmer-2.3.2.bin.intel-linux.tar
(서버 환경에 맞는 hmmer을 다운받는다.(intel, linux, tru64….))

▶ Pfam 릴리즈 파일 다운로드 절차
$ cd ~pfam_home/pfamserver-17.0/pfamdist
$ ftp ftp.genetics.wustl.edu(anonymous)
$ cd pub/Pfam
$ bin
$ mget *
$ bye

6) Installation 프로시저 실행
/pfam_home/pfamserver-17.0/Makefile을 환경에 맞게 수정한후 make pfam 실행

▶ 추가 및 수정내용
* hmmer 파일을 /pfam_home/pfamserver-17.0/bin에 복사하기 위한 내용
(cd bin; cp ../hmmerdist/ hmmer-2.3.2.bin.intel-linux /binaries/* .)
cd /hmmerdist/ hmmer-2.3.2.bin.intel-linux/squid
cp afetch alistat seqstat sfetch shuffle sindex sreformat ../../../bin

```
pfam 예제:
        @echo Installation Proecssing
        (cd bin; cp ../hmmerdist/hmmer-2.3.2.bin.alpha-tru64/binaries/* .)
        cd   /hmmerdist/hmmer-2.3.2.bin.alpha-tru64/squid
        cp afetch alistat seqstat sfetch shuffle sindex sreformat ../../../bin
        mkdir Desc Full Seed Otherdata docs/browse
        @echo Building the description file directory, Desc/...
        (cd Desc; gunzip -c ../pfamdist/Pfam-A.full.gz | ../setup/pfam2desc.pl)
        sort -f -o INDEX INDEX
        mv INDEX conf/
        @echo Building the full alignments directory, Full/...
        (cd Full; gunzip -c ../pfamdist/Pfam-A.full.gz | ../setup/pfam2slx.pl)
        @echo Building the seed alignments directory, Seed/...
        (cd Seed; gunzip -c ../pfamdist/Pfam-A.seed.gz | ../setup/pfam2slx.pl)
        @echo Building binary HMM databases in Otherdata/...
        (cd pfamdist; gunzip -c Pfam_fs.gz > Pfam_fs)
        (cd pfamdist; gunzip -c Pfam_ls.gz > Pfam_ls)
        (cd pfamdist; ../bin/hmmconvert -b Pfam_fs ../Otherdata/Pfam_fs17.0)
        (cd pfamdist; ../bin/hmmconvert -b Pfam_ls ../Otherdata/Pfam_ls17.0)
        (cd Otherdata; ../bin/hmmindex Pfam_fs17.0)
        (cd Otherdata; ../bin/hmmindex Pfam_ls17.0)
        gunzip -c pfamdist/Pfam-B.gz      > Otherdata/Pfam-B
        gunzip -c pfamdist/swisspfam.gz > Otherdata/swisspfam
        gunzip -c pfamdist/pfamseq.gz     > Otherdata/pfamseq
        gunzip -c pfamdist/domain.pnh.gz     > Otherdata/domain.pnh
        (cd Otherdata; ../bin/sindex --pfam pfamseq)
        (cd Otherdata; ../setup/pfamb2gsi.pl Pfam-B)
        (cd Otherdata; ../setup/swisspfam2gsi.pl swisspfam)
        (cd setup; perl makebrowse.pl)
        (cd setup; perl makessi.pl)
```

```
$ make pfam >& log.txt
```

7) 웹서버 동작관련 디렉토리 생성

웹서버가 디렉토리를 사용할 수 있도록 해야 하므로 root 권한으로 디렉토리 생성 작업을 수행한다.

$ cd ~pfam_home/pfamserver-17.0
$ mkdir tmp
$ chmod 755 tmp
$ chown apache tmp

$ mkdir logs
$ chmod 755 logs
$ chown apache logs
$ mkdir docs/images/tmp
$ chmod 755 docs/images/tmp
$ chown apache docs/images/tmp

8) 기타(설치시 고려사항)

▶ pfam설치를 위해서 pfamserver-17.0에서 제공하는 INSTALL문서와 Makefile문서를 참고하여 설치를 수행한다. INSTALL문서는 전체적인 흐름을 설명하는 문서이지만 핵심 compile에 대한 내용은 Makefile에 포함되어 있다.

▶ LWP:UserAgent 모듈의 유, 무에 따라 에러가 발생한다. 기본값으로 LWP:UserAgent 모듈이 동작하도록 설정되어 있으므로 LWP:UserAgent 모듈이 없을 경우에는 동작하지 않도록 조치를 해야 에러를 피할 수 있다.
/pfam_home/pfamserver-17.0/cgi-bin/getdesc안에 LWP:UserAgent 내용이 존재한다.

▶ pfam 서버에 존재하는 모든 *.shtml(Server-side include HTML)은 상단과 하단부분에는 다음과 같은 set, include 문이 존재하는데 띄워쓰기 오류로 인한 에러가 발생한다. 다음과 같이 수정해야 정상적으로 동작한다. 수정해야 할 shtml이 많으므로 스크립트나 shell을 작성하여 변환하는 것이 효율적이다.
대부분의 shtml 파일은 다음 경로에 존재한다.
/pfam_home/pfamserver-17.0/docs/
/pfam_home/pfamserver-17.0/docs/help
/pfam_home/pfamserver-17.0/docs/browse

Shtml 수정 예제:

```
<상단>
<!--#set var="maintitle"    value="Protein Search" -->
<!--#set var="description" value="Analyze a query sequence using the Pfam HMM database" -->
<!--#include virtual="/ssi_header.shtml" -->
                                    .....
<하단>
<!--#include virtual="/ssi_footer.shtml" -->

                                    <<< 수정후 >>>
<상단>
<!--# set var="maintitle"    value="Protein Search" -->
<!--# set var="description" value="Analyze a query sequence using the Pfam HMM database" -->
<!--# include virtual="/ssi_header.shtml" -->
                                    .....
<하단>
<!--# include virtual="/ssi_footer.shtml" -->
```

3. OCA

OCA는 Weizmann 연구소에서 제공하는 검색도구로서 PDB 데이터베이스에 대한 검색을 사용자의 이용목적에 맞도록 『Simple』, 『FASTA』, 『Additional』등과 같이 검색기능을 세분화하여 제공한다. 또한 질의에 대한 검색결과를 사용자의 필요에 따라 저장할 수 있는 기능과 Rasmol, MAGE, VRML 브라우저를 사용하여 검색된 분자의 3차원 구조를 확인 할 수 있는 기능을 지원하고 있다.

▶ OCA 미러설치 절차

- 적당한 위치에 OCA 계정 또는 디렉토리를 생성한 후 OCA brower와 engine 파일을 다운로드 받아 저장한다.
- OCA 미러 서비스를 위해 아파치(httpd) 설정을 수행한다.
- OCA 환경설정 정보를 나타내는 /oca/oca-bin/oca-local 파일을 환경에 맞게 수정

한다.

- OCA 미러를 위해 필요한 관련 소프트웨어를 설치한다.
- 4번에서 설치한 소프트웨어를 soft links을 잡아준다.
- /oca/oca-bin/testInstallation을 웹브라우저를 사용하여 링크 테스트를 수행한다.
- 마지막으로, /oca/oca-support/tools/oca-update을 수행하여 인덱스를 생성한다.

1) OCA 소스파일 다운로드 하기

ftp주소와 계정은 설치 및 업데이트 작업에 따라 weizmann에서 임시적으로 발급해준다.

DB name	OCA
ftp address	biocourse.weizmann.acil
account	ocamir

다운받은 파일:

OCA_Database.20040308.full.tar.gz

OCA_Database.20040308.update.tar.gz

OCA_Database.20040328.update.tar.gz

OCA_Database.20041116.update.tar.gz

OCA_Engine.20040308.full.tar.gz

OCA_Engine.20040308.update.tar.gz

OCA_Engine.20040328.update.tar.gz

OCA_Engine.20041116.update.tar.gz

2) 압축해제 및 디렉토리 생성하기

$ gunzip -c OCA_Engine.YYYYMMDD.full.tar.gz | tar xf -

$ gunzip -c OCA_Engine.YYYYMMDD.update.tar.gz | tar xf -

정상적으로 압축이 해제되고 나면 다음과 같은 디렉토리 구조를 가지게 된다. 일부 디렉토리는 mkdir 명령을 이용하여 직접 생성해준다.

```
oca/
        oca-support/
                db/
                lpc.csu/
                monomers/
                oca-data/
                oca-disease/
                oca-function/
                oca-head/
```

oca-ligands/
oca-perl/
oca-seqs/
tools/
spelling/
oca-bin/
oca-docs/
img/
LIGIM/

$ cd /oca/oca-support
$ mkdir oca-head
$ mkdir oca-seqs
$ cd /oca/oca-docs
$ mkdir LIGIM

3) 아파치 서버(httpd.conf) 설정 및 수정하기
httpd.conf에서 OCA에 관련 한 내용을 추가한다.
oca.ccbb.re.kr httpd.conf 설정화면

```
<VirtualHost *:80>
    ServerName          oca.ccbb.re.kr
    DocumentRoot        /data1/oca/oca-docs
    Alias           /oca-docs/   "/data1/oca/oca-docs/"
    Alias           /img/        "/data1/oca/oca-docs/img/"
    Alias           /oca-ligands/       "/data1/oca/oca-support/oca-ligands/"
    Alias           /LIGIM/          "/data1/oca/oca-docs/LIGIM/"
    ScriptAlias     /oca-bin/    "/data1/oca/oca-bin/"
    AddType         chemical/x-pdb  pdb
</VirtualHost>
```

설정을 마치고 나면 아파치 서버를 restart 한다.

▶ oca-local 파일 수정
oca-local 파일은 OCA 미러 서비스에 관련된 환경정보를 가지고 있다. 이를 시스템에 맞도록 설정하는 작업이 필요하다. oca-local.tmpl 샘플 파일을 oca-local 파일로 복사한 후 작업을 진행한다.
$ cd oca/oca-bin
$ cp oca-local.tmpl oca-local
$ vi oca-local

[oca-local 설정화면]

```
$VERSION = "1.06"; $AUTHOR = "Prilusky"; $YEAR = "1999-2002";

package OCA;
# oca's working directories
   $OCA             = "/data1/oca";
   $support         = "$OCA/oca-support";
   $spelling        = "$OCA::support/spelling";
   $cachedir        = "$OCA/publicTmp";
## By OCS    $cachedir        = "/publicTmp";

# original data locations
   $ftpDir          = "/data1/oca/pdb.rcsb/data";
   $monomers        = "$ftpDir/monomers";
   $divided         = "$ftpDir/structures/divided";
   $sfDir           = "$ftpDir/structures/all/structure_factors";
   $nmr_restraints  = "$ftpDir/structures/all/nmr_restraints";
   $fastaData       = "$OCA::support/db/pdb_seqres.txt";
   $models          = "$ftpDir/structures/models";

# support
   $baseRef         = "http://oca.ccbb.re.kr";
   $dbserver        = "$baseRef/oca-bin";
   $glimpse         = "$support/tools/glimpse";
   $fastaDir        = "$support/tools";
   $data            = "$support/oca-data";
   $ligands         = "$support/oca-ligands";
   $pdbhome         = "http://pdb.weizmann.ac.il";

#  CSU & LPC
#    $docsServer      = "/devel/ocadev-apache/htdocs/";
   $docsServer      = "/data1/oca/oca-docs/";
   $cofbinpath      = "$support/lpc.csu";
   $cofd            = "$docsServer/LIGIM";

# geographical information
   $locale = "CCBB" unless $locale;       # insert here your site codename
   $localArea = 4;          # see $OCA::support/db/oca-mirrors for values

require "$OCA::support/db/oca-links";

# ********** OVERRIDE AREA **********
# If needed, define here new values for variables defined in oca-links
# for example:
# $scopServer = "http://scop.mrc-lmb.cam.ac.uk/scop";
```

4) OCA 서비스 관련 파일 설치

http://bip.weizmann.ac.il/oca-docs/requirments.html을 참조하여 시스템에 설치되어 있지 않은 소프트웨어는 다운로드 받은 후 설치한다.

Linux	
Apache (http server)	http://www.apache.org/dist/httpd/binaries/linux/httpd-2.0.48-x86_64-unknown-linux-gnu.tar.gz
Perl	http://www.activestate.com/Products/Download/Register.plex?id=ActivePerl
fasta3	ftp://ftp.virginia.edu/pub/fasta/fasta3.shar.Z
glimpse (glimpse, glimpseindex, agrep)	http://webglimpse.net/trial/glimpse-latest.tar.gz
gunzip (gzip, gunzip)	http://www.gzip.org
ispell (ispell, buildhash)	http://fmg-www.cs.ucla.edu/geoff/tars

5) soft links 생성

4번 작업까지 완료되고 나면 ln 명령어를 이용하여 soft .links 생성해주는 작업이 필요하다.

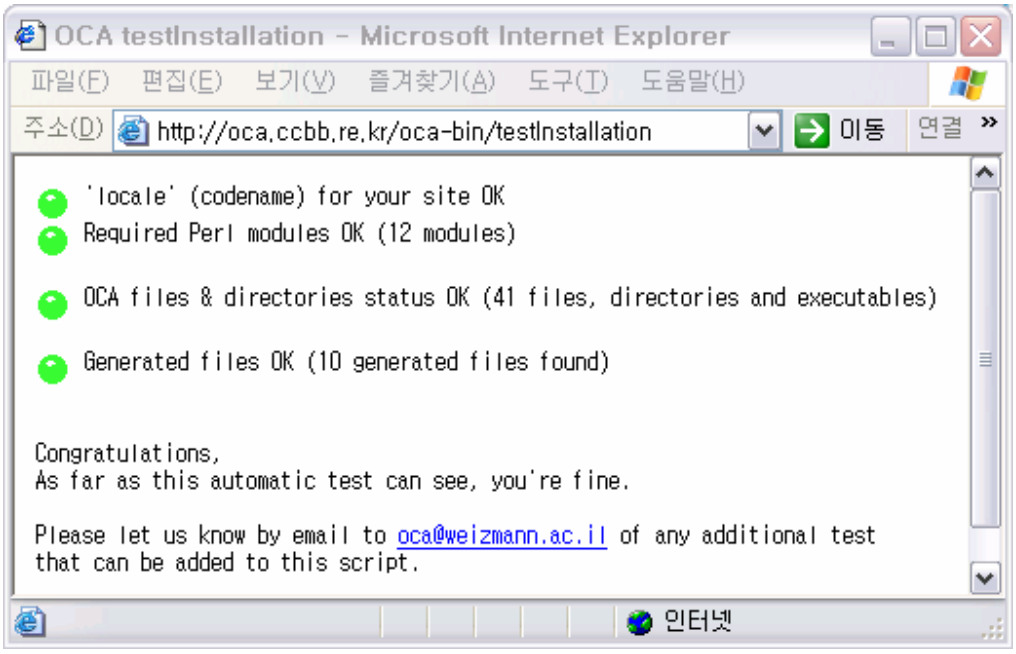
$ cd ~oca/oca-bin
$ ln -s /usr/local/bin/perl perl
$ cd ~oca/oca-support/tools
$ ln -s ~oca/oca-bin/oca-local oca-local
$ ln -s /usr/local/bin/perl perl
$ ln -s /usr/local/bin/agrep agrep
$ ln -s /usr/local/bin/glimpse glimpse
$ ln -s /usr/local/bin/glimpseindex glimpseindex
$ ln -s /usr/local/bin/fasta3 fasta
$ ln -s /usr/local/bin/ispell ispell
$ ln -s /usr/local/bin/buildhash buildhash
$ ln -s /usr/local/bin/gzip gzip
$ ln -s /usr/local/bin/gunzip gunzip

ftp.rcsb.org에서 다운로드 받은 pdb 데이터 중에서 pdb_seqres.txt에 대한 soft link도 생성한다.
$ ln -s ~oca/pdb.rcsb/pdb_seqres.txt pdb_seqres.txt

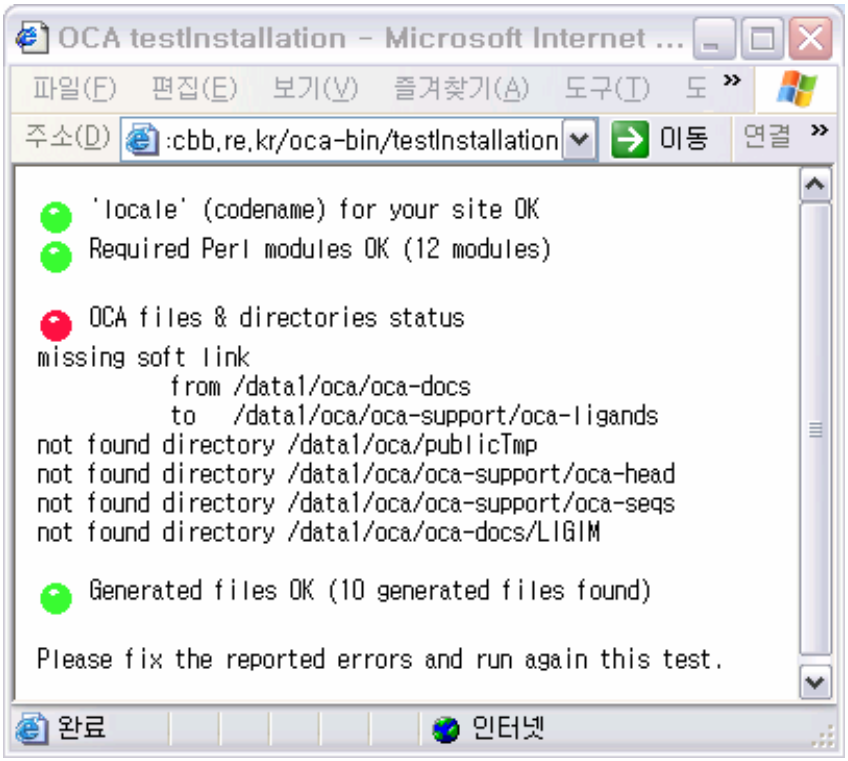
6) make 작업 수행
시스템 OS에 해당 하는 make 파일을 실행한다.
$ cd ~oca/oca-support/lpc.csu/
$ ./makeSUN

7) 설치 결과 테스트
oca-bin/testInstallation 스크립트를 웹브라우저에서 실행하여 1-6번까지 작업이 잘 수행되었는지를 체크한다.

[정상적으로 설치가 성공한 경우]



[비정상적인 경우]



필요한 디렉토리가 생성되지 않은 경우 또는 링크가 없는 경우에 에러 메시지를 출력한다. 이를 참조하여 문제를 해결할 수 있다.

8) 참조메일

설치작업을 수행하면서 궁금한 점에 대한 개발자로부터 얻은 답변 메일이다.

> question 1) we can't find some files(oca-oneline.txt, oca-dates)?
You may try to run oca-update once and then run testInstallation again.

> question 2) we are just make some directory, is right?
Yes. Create any directory reported as missing and set the mode
for the user running Apache (nobody?) to be able to read it.

> question 3) which a makefile better our system?
Try /oca/oca-support/lpc.csu/makeSUN

> question 4) how to setting chmod about 'LIGIM'?
> testInstallation script error message
> "unable to write into directory /oca/oca-docs/LIGIM

Create a directory somewhere, and set the owner to the user
running the web server (nobody?).
Enter the full path of this directory in oca-local
$cofd = "/usr/local/web-apache/LIGIM";
and create an ALias entry on your Apache config file
Alias /LIGIM /usr/local/web-apache/LIGIM

---

9) pdb 데이터 다운로드

OCA는 pdb 원문 데이터를 기반으로 검색할 수 있는 기능을 제공함으로 같은 서버내에 pdb파일이 존재해야 한다.

oca.ccbb.re.kr에서는 oca 계정아래 pdb.rcsb라는 디렉토리를 생성한 후 PDB 데이터를 다운로드 받았다. pdb 데이터에 대한 정보는 ~oca/oca-bin/oca-local 파일을 참조하면 된다.

```
$ cd ~oca/pdb.rcsb/
$ ftp ftp.rcsb.org
$ cd pub/pdb/
$ bin
$ get -R *
```

KiSTi www.kisti.re.kr
한국과학기술정보연구원
Korea Institute of Science and Technology Information

대전 본원
대전광역시 유성구 어은동 52번지

서울 분원
서울특별시 동대문구 청량리동 206-9

ISBN 89-5884-458-2 93560