Unifying Bio-information on the Grid

Hideo Matsuda
Department of Bioinformatic Engineering
Osaka University

E-mail: matsuda@ist.osaka-u.ac.jp

Data Grid for Bio-information

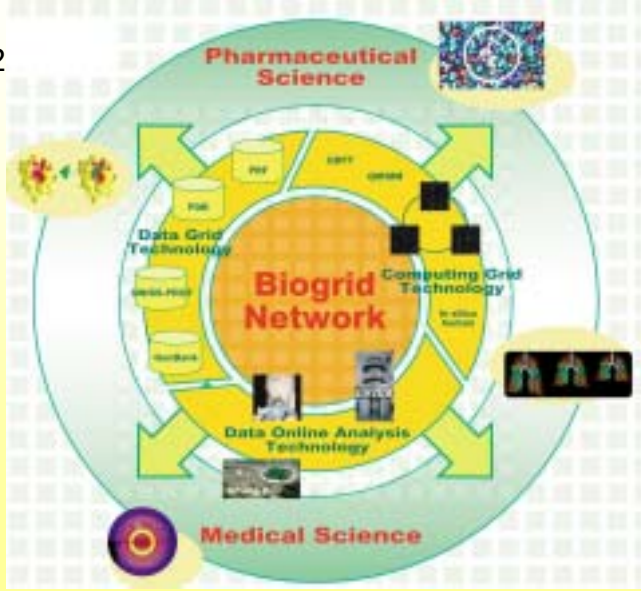
- Bio-information has tremendous amount data and it's still rapidly increase.
- Distributed into many databases (500+ DBs)
  - Very heterogeneous (different contents, different formats)
  - Many similar and related data (need approximate matching, difficult to be indexed)
  - Physical integration is not easy
- ⇨ Coordinating DBs by the Grid

Our BioGrid

http://www.biogrid.jp

- Started from 2002
- Goals

  Technology Development for: Pharmaceutical (drug discovery) and Medical Sciences.
- Leader: Shinji Shimojo (CMC, Osaka Univ.)
- Government Support (MEXT): 5years (3~4M$/year)



Heterogenity of Bio-Databases (Flat Files)

DNA sequence DB(DDBJ)

```
LOCUS       ECORBS 6197 bp DNA
            30-OCT-1994
DEFINITION  E.coli ...rbsB ....
            transport system, ..
ACCESSION   M13169 M13517
NID         g147511
KEYWORDS    high affinity ribose
SOURCE      E.coli K12 DNA.
ORGANISM    Escherichia coli
REFERENCE   1  (sites)
  AUTHORS   ...
  TITLE     ...
  JOURNAL   ...
  MEDLINE   84032513
MEDLINE     84032513
FEATURES
CDS             3112..4002
     /gene="rbsB"
     /translation="MKKGTVLNS...
ORIGIN      94 bp upstream of BclI
            1 ctcaggttcg aaatctaac
```

Protein sequence DB (SWISS-PROT)

```
ID   RBSB_ECOLI ...
AC   P02925;
DT   21-JUL-1986 (...CREATED)
DT   21-JUL-1986 (... UPDATE)
DT   01-NOV-1995 (... UPDATE)
DE   D-RIBOSE-BINDING ...
GN   RBSB OR RBSP OR PRLB.
OS   ESCHERICHIA COLI.
OC   PROKARYOTA; GRACILICUTES;
OC   ENTEROBACTERIACEAE.
RN   [1]RP SEQUENCE FROM N.A.,...
RX   MEDLINE; 84032513.
RA   GROARKE J.M., MAHONEY W.C.,
RA   ZALKIN H., HERMODSON M.A.;
RL   J. BIOL. CHEM. 258:...
CC   -!- FUNCTION: INVOLVED IN
 ...
KW   TRANSPORT; SUGAR TRANSPORT;
KW   3D-STRUCTURE.
SQ   SEQUENCE   296 AA; ...
     MNMKKLATLV SAVALSATVS ANA...
     ...
```

Protein 3D structure DB (PDB)

```
HEADER    SUGAR TRANSPORT 23-SEP-9
COMPND    D-RIBOSE-BINDING PROTEIN
COMPND    2 (G134R) COMPLEXED WITH
SOURCE    (ESCHERICHIA COLI)...
SOURCE    2 EXPRESSION PLASMID)
AUTHOR    S.L.MOWBRAY,A.J.BJORKMAN
REVDAT    1   26-JAN-95 1DRJ    0
JRNL        AUTH   A.J.BJORKMAN
JRNL        TITL   PROBING PR...
JRNL        TITL 2 RIBOSE-BINDING
JRNL        REF    TO BE PUBLISHED
JRNL        REFN   ASTM
SEQRES  1  271  LYS ASP THR ...
SEQRES  2  271  PRO PHE PHE ...
...
HELIX   1   A PRO     14  LEU ...
HELIX   2   B PRO     43  LEU ...
...
ATOM      1  N   LYS   1  x1 y1 z1
ATOM      2  CA  LYS   1  x2 y2 z2
...
```

BioDBs now provide XML formats

- But they are still heterogeneous!

```
LOCUS       ECORBS 6197 bp DNA ..
            30-OCT-1994
DEFINITION  E.coli ...rbsB ....
            transport system, ...
ACCESSION   M13169
KEYWORDS    high affinity ribose ..
SOURCE      E.coli K12 DNA.
ORGANISM    Escherichia coli
REFERENCE   1  (sites)
  AUTHORS   ...
  TITLE     ...
  JOURNAL   ...
  MEDLINE   84032513
FEATURES
CDS             3112..4002
     /gene="rbsB"
     /translation="MKKGTVLNS...
                    ...
ORIGIN   94 bp upstream of BclI
         1 ctcaggttcg aaatctaac. ...
```

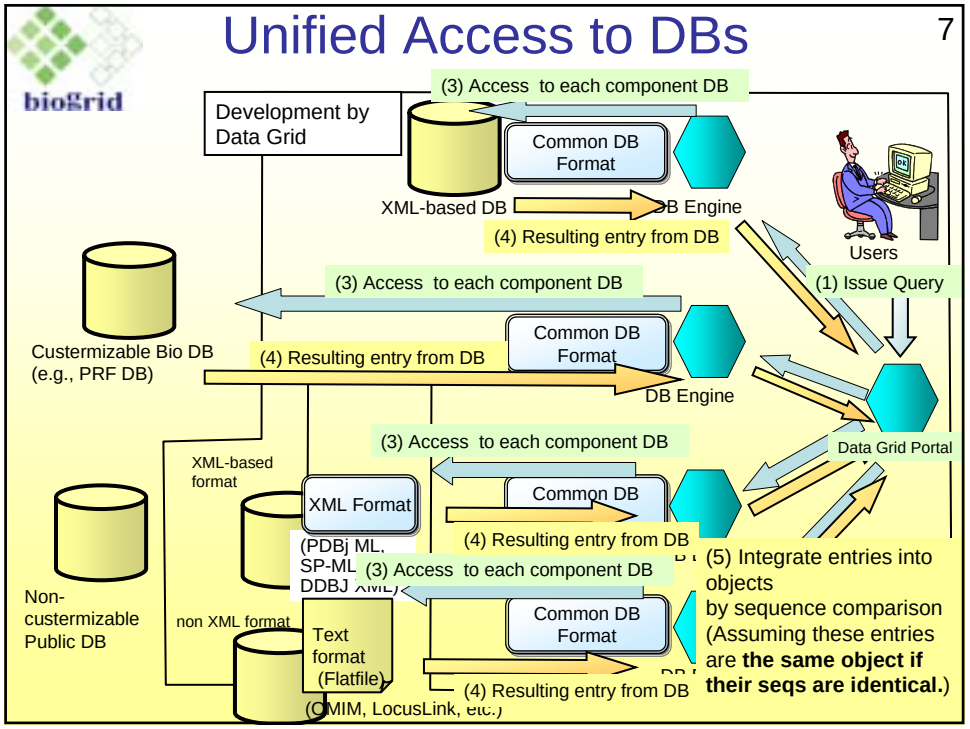
DDBJ FF to DDBJ XML

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE DDBJXML SYSTEM "DDBJXML.dtd">
<DDBJXML>
<LOCUS> ECORBS </LOCUS>
    <LAST_UPDATE> 30-OCT-1994 </LAST_UPDATE>
<DEFINITION> E.coli ...rbsB ....
          transport system, ...</DEFINITION>
<ACCESSION>  M13169 </ACCESSION>
<KEYWORDS> high affinity ribose ..</KEYWORDS>
<SOURCE> Escherichia coli </SOURCE>
<ORGANISM> Escherichia coli </ORGANISM>
<REFERENCE ID="1">
  <AUTHORS>   ...  </AUTHORS>
  <TITLE>     ...  </TITLE>
  <JOURNAL>   ... </JOURNAL>
<FEATURES>
<CDS>     <location>3112..4002</location>
     <qualifiers name="gene"> rbsB </gene>
     <qualifiers name="translation"> MKKGTVLNS...
     </qualifiers>
                    ...
<SEQUENCE> ctcaggttcgaaatctaac. ...</SEQUENCE>
```


Common Data Translation Table

	Common Format	SWISS-PROT	PIR	PDB
Name space prefix	---	sp	pir	pdb
Entry (1 protein)	entry	entry	ProteinEntry	PDBj
Entry ID	entry/id	entry/@name	ProteinEntry/@id	PDBj/@entry_ID
Accession	entry/accession	entry/@primaryACC	ProteinEntry/header/accession	PDBj/@entry_ID
Protein name	entry/name	entry/proteinName	ProteinEntry/protein/alt-name	PDBj/main/entity/entity_item/desc iption
Protein alt name	entry/alt-name	entry/proteinName/synonyms/syno nym	ProteinEntry/protein/name	PDBj/main/entity/entity_item/com name
Gene name	entry/gene	entry/genes/gene/@name	ProteinEntry/genetics/gene/uid	PDBj/main/entity/entity_item/src_ en/gene
Species (formal)	entry/organism/scientif ic	entry/organisms/organism/name	ProteinEntry/organism/formal	PDBj/main/entity/entity_item/src_ en/scientific_name
Species (common)	entry/organism/commo n	entry/organisms/organism/name	ProteinEntry/organism/common	PDBj/main/entity/entity_item/src_ en/common_name
NCBI Taxonomy ID	entry/organism/taxId	entry/organisms/organism/taxId	---	---
Reference	entry/reference	entry/references/reference	ProteinEntry/reference/refinfo	PDBj/main/citation/citation_item
Reference ID (MELINE)	entry/reference/medlin eID	entry/references/reference/pubMe d/@medlineID	ProteinEntry/reference/refinfo/xref s/xref/uid[db="MUID"]	---
Reference ID (PubMed)	entry/reference/pubme dID	entry/references/reference/pubMe d/@value	ProteinEntry/reference/refinfo/xref s/xref/uid [db="PMID"]	---
Function	entry/function	entry/comments	---	---
Enzyme EC#	entry/EC_num	entry/proteinName	---	PDBj/main/entity/entity_item/EC
GeneOntology	entry/GO	entry/goTerms	---	---
GO term	entry/GO/term			
GO description	entry/GO/description			
Keyword	entry/keyword	entry/keywords/keyword	ProteinEntry/keywords/keyword	PDBj/main/struct/keywords
Structure	entry/structure			PDBj/main/struct/struct
Sequence (amino acid)	entry/sequence	entry/sequence	ProteinEntry/sequence	PDBj/main/entity/entity_item/seq_ one_letter_code
Feature	entry/feature	entry/features/feature	ProteinEntry/feature	PDBj/struct/site
feature type	entry/feature/type	entry/features/feature/@key	ProteinEntry/feature/feature-type	PDBj/struct/site/@id
feature description	entry/feature/descriptio n	entry/features/feature/@descriptio n	ProteinEntry/feature/description	PDBj/struct/site/site_gen/details

Unified Access to DBs

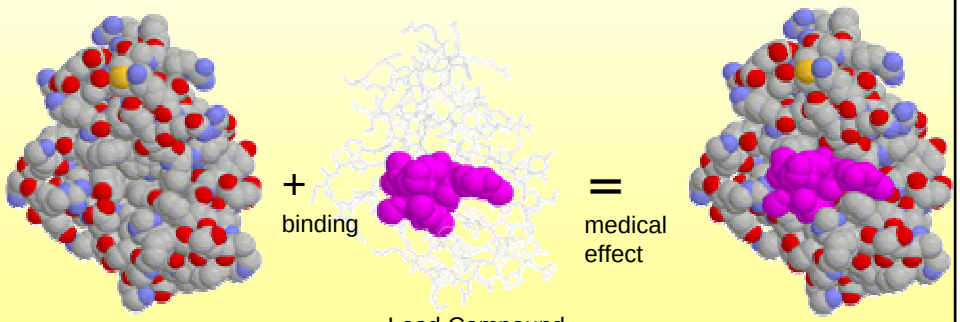


An Example of Common Database Format


```
<?xml version="1.0"?>
<entry xmlns="bio data grid" xmlns:sp="swiss-prot" xmlns:pir="pir" xmlns:pdb="pdb" >
  <id>..</id> <acc>..</acc> <date> .. </date>
  <name db="pir" acc=".." >protein</name>  <alt-name db="pir" acc="..">..</alt-name>
  <gene db="sp" acc="..">gene name</gene>
  <organism db="pir" acc="..">
    <scientific db="pir" acc="..">Homo sapiens</scientific>
    <common db="pir" acc="..">human</common> <taxId db="pir" acc="..">..</taxId>
  </organism>
  <reference db="pir" acc="..">
    <author>..</author> <citation>Journal of ..</citation>  <volume>1</volume>
    <year>2002</year>
    <title>paper title</title> <first_page>10</first_page> <last_page>20</last_page>
    <medlineID>..</medlineID> <pubmedID>..</pubmedID>
  </reference>
  <function db="sp" acc=".."> <desc>..</desc> </function>
  <keyword db="sp" acc="..">..</keyword>
  <GO db="sp" acc=".."> <term>0006118</term><desc>electron trans..</desc> </GO>
  <feature db="pir" acc=".."> <type>active site</type> <desc>..</desc>
    <start>..</start> <end>..</end> </feature>
  <sequence db="sp" acc=".." length=".."> MDPVVVLVLGL.. </sequence>
  <sp:data>SWISS-PROT data</sp:data>
  <pir:data>PIR data</pir:data>
  <pdb:data>PDB data</pdb:data>
</entry>
```

Application of Data Grid to Drug Discovery

- Drug Related Data:
  - Drug Target Proteins (mainly receptors)
  - Lead Compounds (drug candidates)
  - Medical Effect
- Correlate all information by unified DB access (active site locations, charges, known binding partners, etc.)



Target Protein (e.g., Immunophilin FKBP12)

Lead Compound (e.g., FK506)

Immuno-suppression

Summary

- Biology is now going to be an *Information Science*.
- Bio DBs are very heterogeneous. We need some unified access with some *common* data format.
- Coordinating DBs are useful for biological applications: Drug Discovery, etc.