

Web services and genome annotation in GRID by DNA Data Bank of Japan (DDBJ)

Center for Information Biology and DNA Data Bank of Japan

National Institute of Genetics

Hideaki Sugawara and Satoru Miyazaki

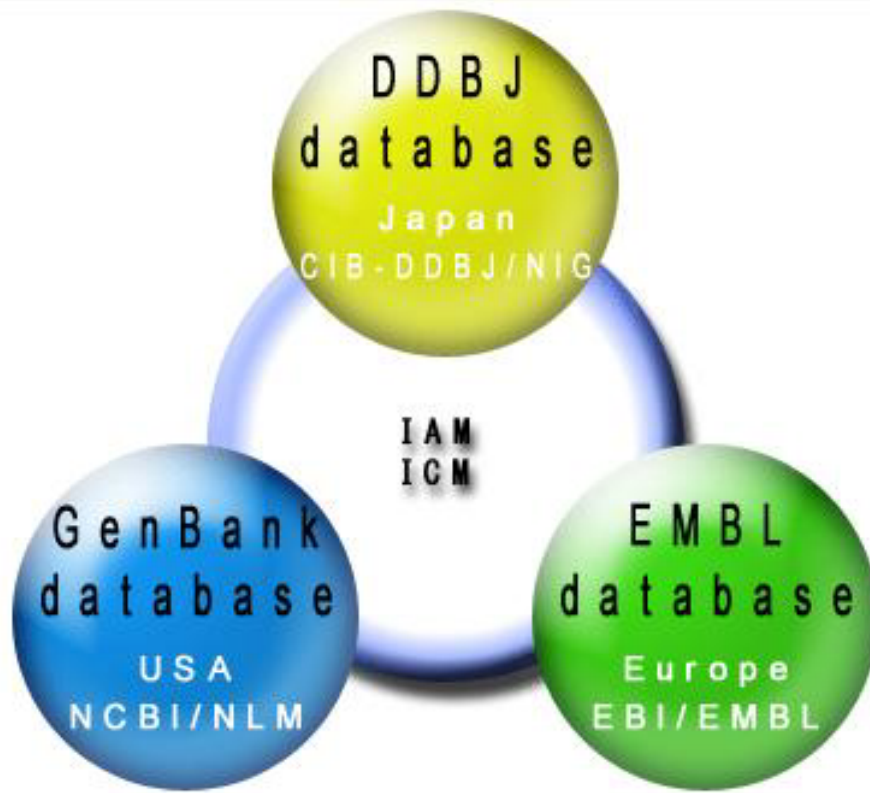
hsugawar@genes.nig.ac.jp

smiyazak@genes.nig.ac.jp

Contents

- Background and motivation
- SOAP servers by DDBJ
- Web services by DDBJ
- A work flow
- GRID test-bed

The International Nucleotide Sequence Database (INSD): DDBJ/EMBL/GenBank



DDBJ/EMBL/GenBank
International
Nucleotide Sequence Database

DDBJ: DNA Data Bank of Japan
CIB-DDBJ: Center for Information Biology and DNA Data Bank of Japan
NIG: National Institute of Genetics

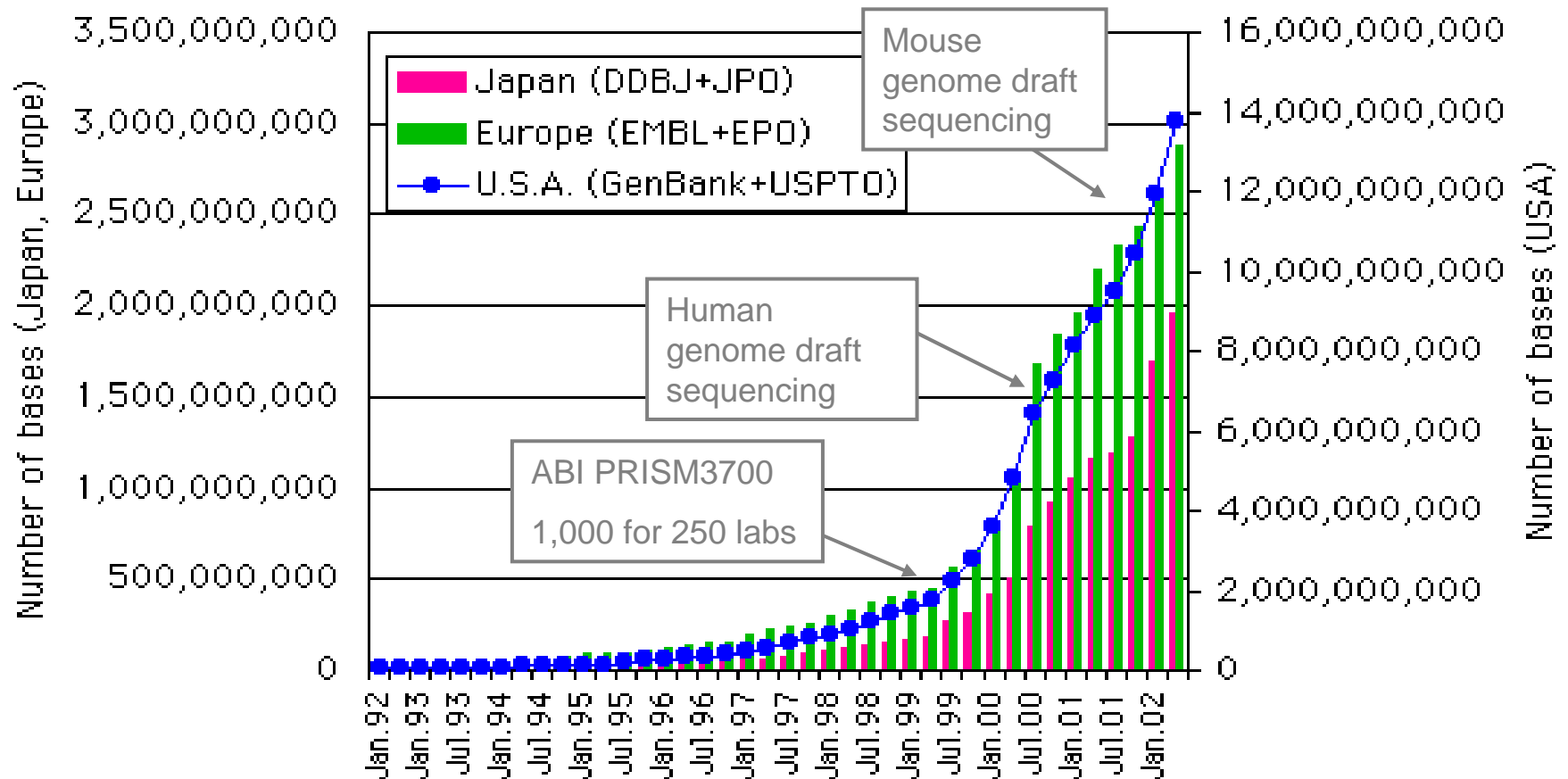
EMBL: European Molecular Biology Laboratory
EBI: European Bioinformatics Institute

NCBI: National Center for Biotechnology Information
NLM: National Library of Medicine

IAM: International Advisory Meeting
ICM: International Collaborative Meeting

Number of bases (atgc) in INSD

Contribution of the three geographical areas to the DDBJ/EMBL/GenBank International Nucleotides Database (1992/1-2002/4)



Genome projects and biodiversity studies are going on

	US	Europe	Japan	Others
Archaea	11	4	2	1
Procaryote	168	63	12	14
Eucaryote	112	54	6	3

Ref <http://wit.integratedgenomics.com/GOLD/>

Environmental sequences in INSD	
2002/09	97,512 entries
2002/11	107,936 entries
2003/02	149,284 entries

Nucleic Acids Research

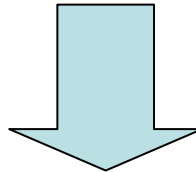
Database Categories List

www3.oup.co.uk/nar/database/c/

- ▶ Major Sequence Repositories
- ▶ Comparative Genomics
- ▶ Gene Expression
- ▶ Gene Identification and Structure
- ▶ Genetic and Physical Maps
- ▶ Genomic Databases
- ▶ Intermolecular Interactions
- ▶ Metabolic Pathways and Cellular Regulation
- ▶ Mutation Databases
- ▶ Pathology
- ▶ Protein Databases
- ▶ Protein Sequence Motifs
- ▶ Proteome Resources
- ▶ RNA Sequences
- ▶ Retrieval Systems and Database Structure
- ▶ Structure
- ▶ Transgenics
- ▶ Varied Biomedical Content

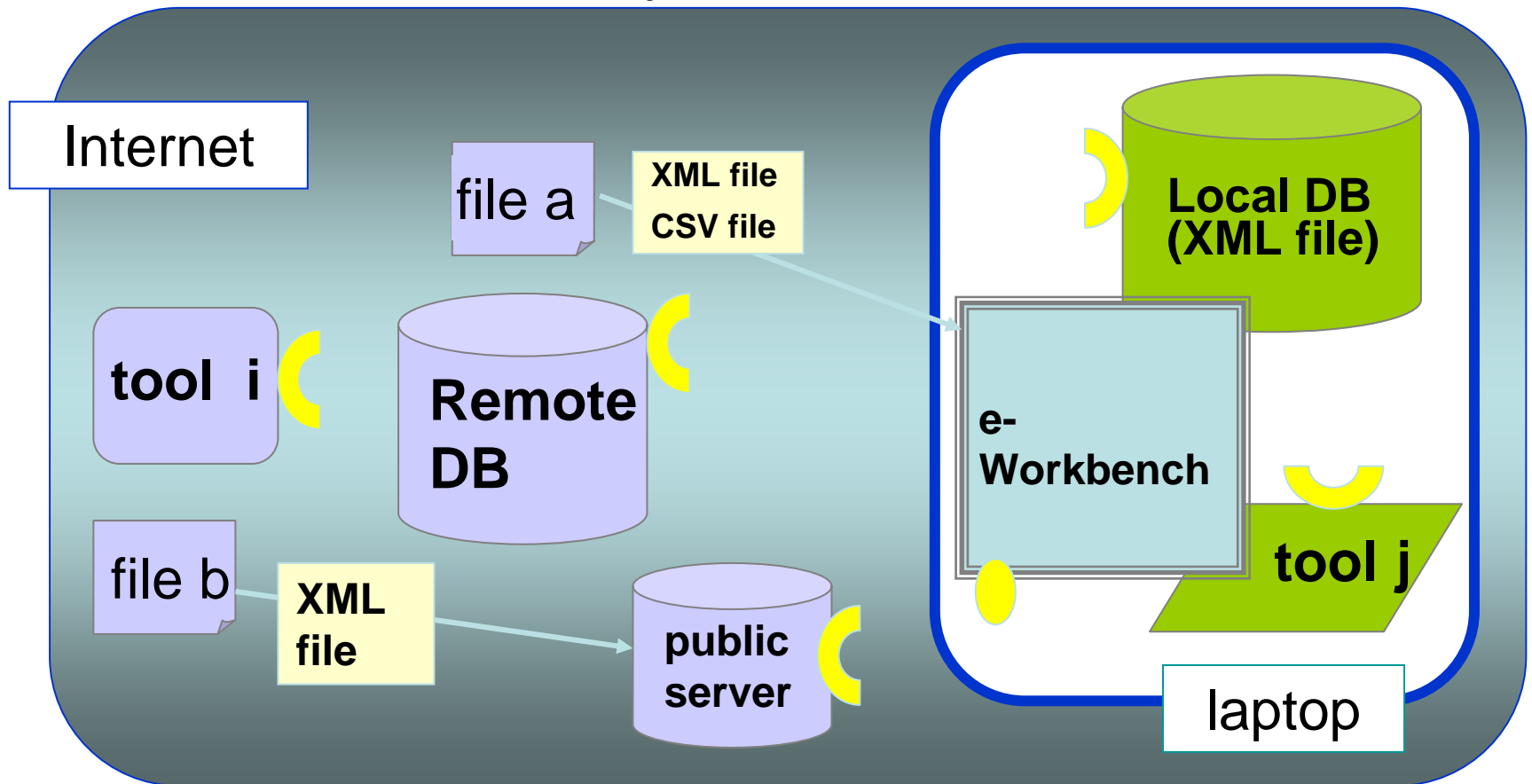
A mission of DDBJ

- Biological data resources are diverse
- Some biological data resources are very large scale databases (VLSD)
- Diverse requirements to integrate these biological data resources



- To contribute to the interoperability of biological data resources

Integration of distributed diverse data sources by use of CORBA & XML



: CORBA (Common Object Request Broker Architecture)

DDBJ-XML

DDBJ entry

LOCUS AK025000 2589 bp mRNA HUM 29-SEP-2000
 DEFINITION Homo sapiens cDNA: FLJ21347 fis, clone COL02724.
 ACCESSION AK025000
 VERSION AK025000.1
 KEYWORDS oligo capping; fis (full insert sequence).
 SOURCE Homo sapiens colon cDNA to mRNA, clone_lib:COL clone:COL02724.
 ORGANISM Homo sapiens
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
 REFERENCE 1 (bases 1 to 2589)
 AUTHORS Sugano,S., Suzuki,Y., Ota,T., Obayashi,M., Nishi,T., Isogai,T.,
 Shibahara,T., Tanaka,T. and Nakamura,Y.
 TITLE Direct Submission
 JOURNAL Submitted (29-AUG-2000) to the DDBJ/EMBL/GenBank databases. Sumio
 Sugano, Institute of Medical Science, University of Tokyo,
 Laboratory of Genome Structure

FEATURES
source

CDS

Location/Qualifiers
 1..2589
 /clone="COL02724"
 /clone_lib="COL"
 /note="cloning vector pME18SFL3"
 /organism="Homo sapiens"
 /sequenced_mol="cDNA to mRNA"
 /tissue_type="colon"
 18..2378
 /codon_start=1
 /protein_id="BAB15051.1"
 /translation="MLGARAWLGRVLLLPRAGAGLAASRRGSSSRDKDRSATVSSSV

annotation

protein sequence

DNA sequence

BASE COUNT 529 a 797 c 773 g 490 t 0 others
 ORIGIN
 1 atcgccccg agcagccatg ctgggcgcgc gggcctggtt gggccgcgtc cttctgctgc

The data structure

- The DDBJ/EMBL/GenBank **Feature Table**: Definition
 - **Feature key**: *e.g. CDS (coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein; ---)*
 - **Qualifier**: *e.g. /gene (symbol of the gene corresponding to a sequence region)*
 - **Value**, *e.g. “text”*
- **Taxonomy Database**

- **Taxonomic name**: Escherichia coli
[\[Go to lower taxa\]](#) [\[Get sequence\(s\) related to this taxon\]](#)
- **Full lineage**: [cellular organisms](#); [Bacteria](#); [Proteobacteria](#); [gamma subdivision](#); [Enterobacteriaceae group](#); [Enterobacteriaceae](#); [Escherichia](#)

ex. CDS of DDBJ/EMBL/GenBank AB000100

FF (Flat File) format

```
CDS      121..912
         /gene="cynB"
         /codon_start=1
         /transl_table=11
         /product="intrinsic membrane protein"
         /protein_id="BAA21794.1"
         /translation="MVRTPVPLYLRWAVSILSVLAFLAIWQIAAASGFLGKTFFPGSLR
         TLQDLFGWLSDPFFDNGPNDLGIGWNLLISLRRVAIGYLLATVVAIPLGIAIGMSALA
         -----"
```

XML document

```
<cds>
  <location>121..912</location>
  <qualifiers name="codon_start">1</qualifiers>
  <qualifiers name="gene">cynB</qualifiers>
  <qualifiers name="product">intrinsic membrane protein</qualifiers>
  <qualifiers name="protein_id">BAA21794.1</qualifiers>
  <qualifiers name="translation">MVRTPVPLYLRWAVSILSVLAFLAIWQIAAASGFLGKTFFG
    SLRTLQDL ----- LLDQGFRFLENQFSYAGNR</qualifiers>
  <qualifiers name="transl_table">11</qualifiers>
</cds>
```

DDBJ SOAP

DDBJ SOAP servers

- BLAST (homology search)
- FASTA (homology search)
- SSearch(Smith-Waterman homology search)
- GetEntry (retrieve entries by Acc#s)
- DDBJ (get the DDBJ full entry and extract some Features)
- ClustalW (multiple alignment)
- SRS (Sequence Retrieval System)
- TxSearch (Taxonomy database Search)

DDBJ WSDL



XML Central of DDBJ

“XML Central of DDBJ” has been partly supported by [BIRD](#) of Japan Science and Technology Corporation (JST)

»» What's New

»» DDBJ-XML **XML**

DDBJ-XML is a new output format of DDBJ entries. It is readable both for human and machine.

»» Web services **SOAP**

This is the first public SOAP service for biology in Japan. The project aims at the standardization of bioinformatics services on the Internet and the improvement of the interoperability. This page also provides you a Web interface of the SOAP server.

»» Registration/Publication of your Web services

You are courteously invited to register your Web service(s) in the list, if you open bioinformatics SOAP service(s) to the public.

»» Sample program to bind a web service to your program

You can use a Web service by specifying the URL address, method and parameter (s) in your program such as Java or perl.

»» Demo system using SOAP technology

You can get the seamless access to the retrieval of DDBJ entry and BLAST execution.
Enjoy SOAP technology.

»» Workflow example using web services

A WORK FLOW composed of the Web services

»» SOAP tutorial

It's easy to access the SOAP services.
This is the first step to try web services.

»» Links

Gateway to biological XMLs and retrieval sites for XML documents

Name	URL	Document	Registrant	Methods
BLAST Demo	http://xml.nig.ac.jp/wsdl/BlastDemo.wsdl	document javadoc	XML Central of DDBJ	List to execute
Blast	http://xml.nig.ac.jp/wsdl/Blast.wsdl	document javadoc	XML Central of DDBJ	List to execute
ClustalW	http://xml.nig.ac.jp/wsdl/ClustalW.wsdl	document javadoc	XML Central of DDBJ	List to execute
DDBJ	http://xml.nig.ac.jp/wsdl/DDBJ.wsdl	document javadoc	XML Central of DDBJ	List to execute
Fasta	http://xml.nig.ac.jp/wsdl/Fasta.wsdl	document javadoc	XML Central of DDBJ	List to execute
GetEntry	http://xml.nig.ac.jp/wsdl/GetEntry.wsdl	document javadoc	XML Central of DDBJ	List to execute
RequestManager	http://xml.nig.ac.jp/wsdl/RequestManager.wsdl	document javadoc	XML Central of DDBJ	List to execute
SRS	http://xml.nig.ac.jp/wsdl/SRS.wsdl	document javadoc	XML Central of DDBJ	List to execute
TxSearch	http://xml.nig.ac.jp/wsdl/TxSearch.wsdl	document javadoc	XML Central of DDBJ	List to execute

A list of methods in the Web services named DDBJ

Method

getFFEntry(accession)

getXMLEntry(accession)

getFeatureInfo(accession, feature)

getAllFeatures(accession)

getRelatedFeatures(accession, start, stop)

getRelatedFeaturesSeq(accession, start, stop)

ex.1: Find and list sub-sequences
that are annotated
(features are attached)

DEMO

[Use case]

Retrieve all the sub-sequence with annotation
(features) between 59000th base and 64000th base of
AL121903

[Method]

`getRelatedFeatures(accession, start, stop)`

ex.1: Find and list sub-sequences
that are annotated
(features are attached)

[Result]

repeat_region 423..717

CDS join(37..121,4775..4917)

repeat_region 1775..2064

repeat_region 2067..2362

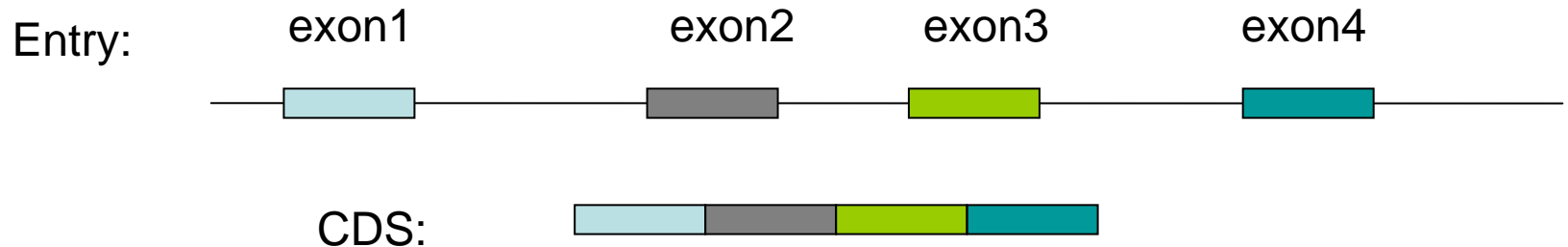
source 1..5001

repeat_region 3067..3374

mRNA join(26..121,4775..4917)

ex.2: Find and list CDSs

[Use case] Retrieve CDSs by concatenating exons



[Method]

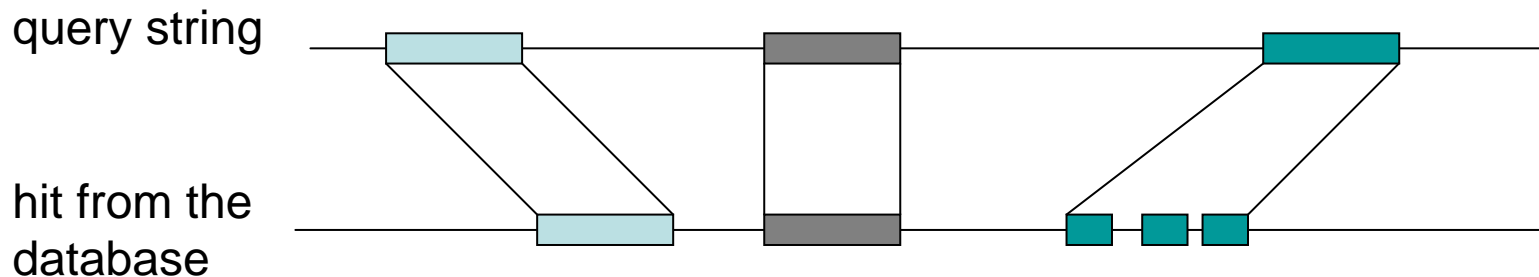
ENDPOINT	<code>http://xml.nig.ac.jp/xddb/GetEntry</code>
METHOD	<code>getFASTA_CDSEntry</code>
INPUTS	<p>accession</p> <input type="text" value="L17418"/>
OUTPUTS	Result <string>

ex.3: Find and list the aligned sub-sequences in the result of blast

[Use case]

Make a file of coordinates of sub-sequences from the result of blast

[Method] `extractPosition(result)`



ex.3: Find and list the aligned sub-sequences in the result of blast

[Result]

AF058428 | AF058428.1

Query	86	248
-------	----	-----

Hit	86	248
-----	----	-----

Query	320	384
-------	-----	-----

Hit	320	384
-----	-----	-----

Query	564	601
-------	-----	-----

Hit	561	598
-----	-----	-----

• • •

ex.4: Find the full lineage

DEMO

[Use case] Check/retrieve the lineage information of species

Lineage Search system

This page is a batch system for searching related lineage information from multiple

Select rank which you want to include the list.

All

Extract ranks

superkingdom phylum class order family

Input your genera (& species) names with format of a name per line.
e.g.
Campylobacter coli
Escherichia coli

Campylobacter coli
Escherichia coli

Search result

download

Query	class	order
Campylobacter coli	Epsilonproteobacteria	Campylobacterales
Escherichia coli	Gammaproteobacteria	Enterobacteriales

[return search page](#)

Tutorial 1: Understand the usage

ENDPOINT `http://oak.genes.nig.ac.jp/glue/urn:blast`

METHOD `search`

INPUTS

program

`<string>`

database

`<string>`

query

`<string>`

param

`<string>`

OUTPUTS `Result <string>`

Execute

Tutorial 2: Understand the function

BLASTS Demo

This is the DDBJ SOAP Demo system.

Enter the Accession Number and you can get three BLAST results.

-> blastn(DDBJ Bacteria)

GetEntry -> blastx(SwissProt) -> result

-> blastx(PDB)

Accession Number: (e.g. AF058429)

Retrieve data from the nucleotide sequence database (INSD), the protein sequence database (SWISS-PROT) and the protein 3D structure database (PDB) all together by an accession number (Acc#) referred in a published paper

Simplified registry

WSDL List

[WSDL Registrant Menu](#)

Name	URL	Document	Registrant	Execute
GetEntry2	http://xml.nig.ac.jp/wsdl/getentry.wsdl	document javadoc	Yasumasa Shigemoto2	<input type="button" value="Execute"/>

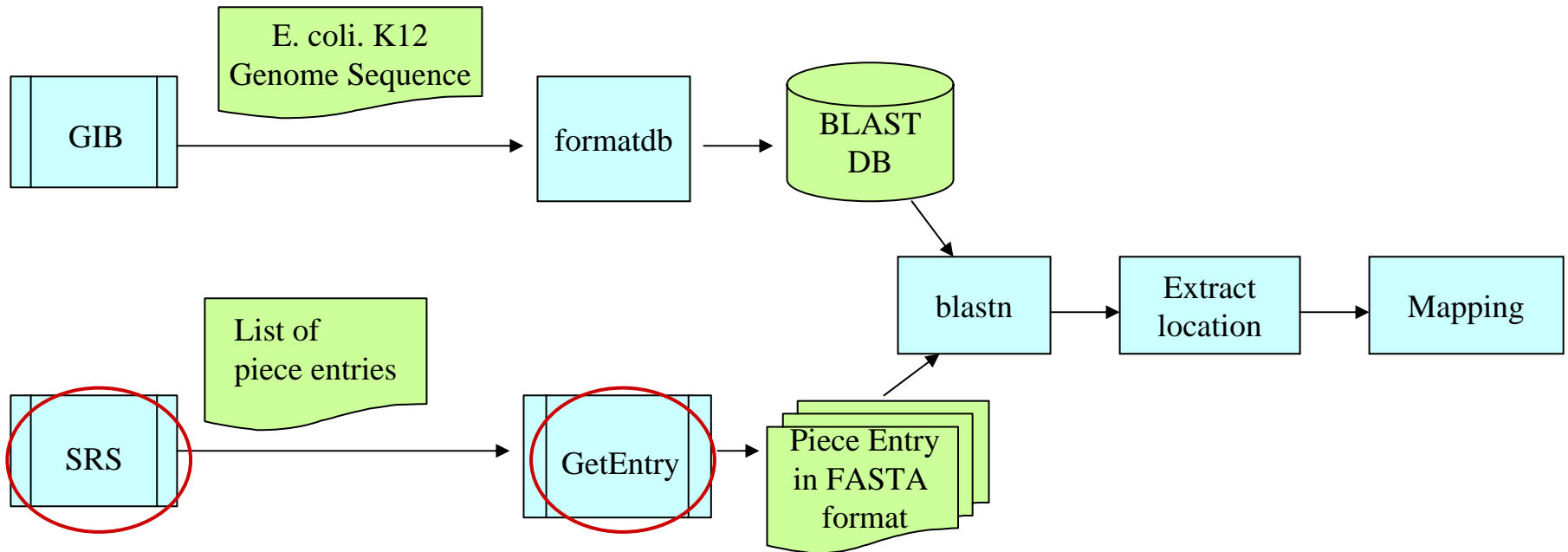
WSDL Update

Site Name *	<input type="text" value="GetEntry"/>
WSDL URL *	<input type="text" value="http://xml.nig.ac.jp/wsdl/getentry.wsdl"/>
Document URL	<input type="text" value="http://xml.nig.ac.jp/doc/GetEntry.txt"/>
JavaDoc URL	<input type="text" value="http://xml.nig.ac.jp/javadoc/GetEntry.html"/>

Work Flow

Map piece entries to genome

Example: Escherichia coli K12 MG1655

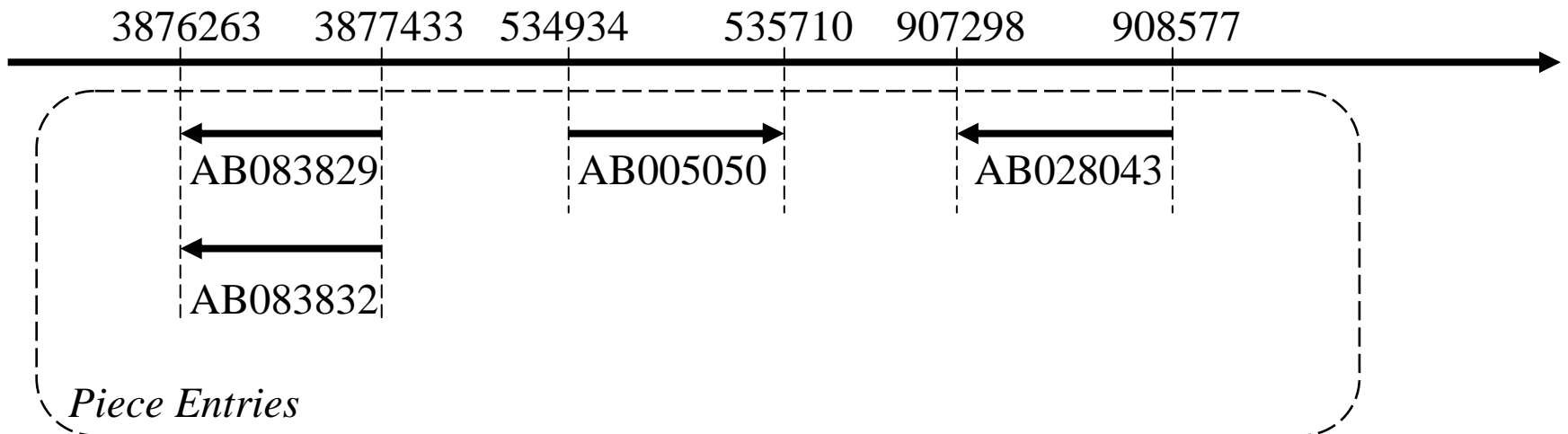


The result of the work flow

3,292 piece entries are mapped to genome sequence

Accession	Start	Stop
AB005050	908577	907298
AB028043	534934	535710

Genome (Eshcherichia coli K12 MG1655)



GRID Use case: an annotation project

ORF detection

Clustering

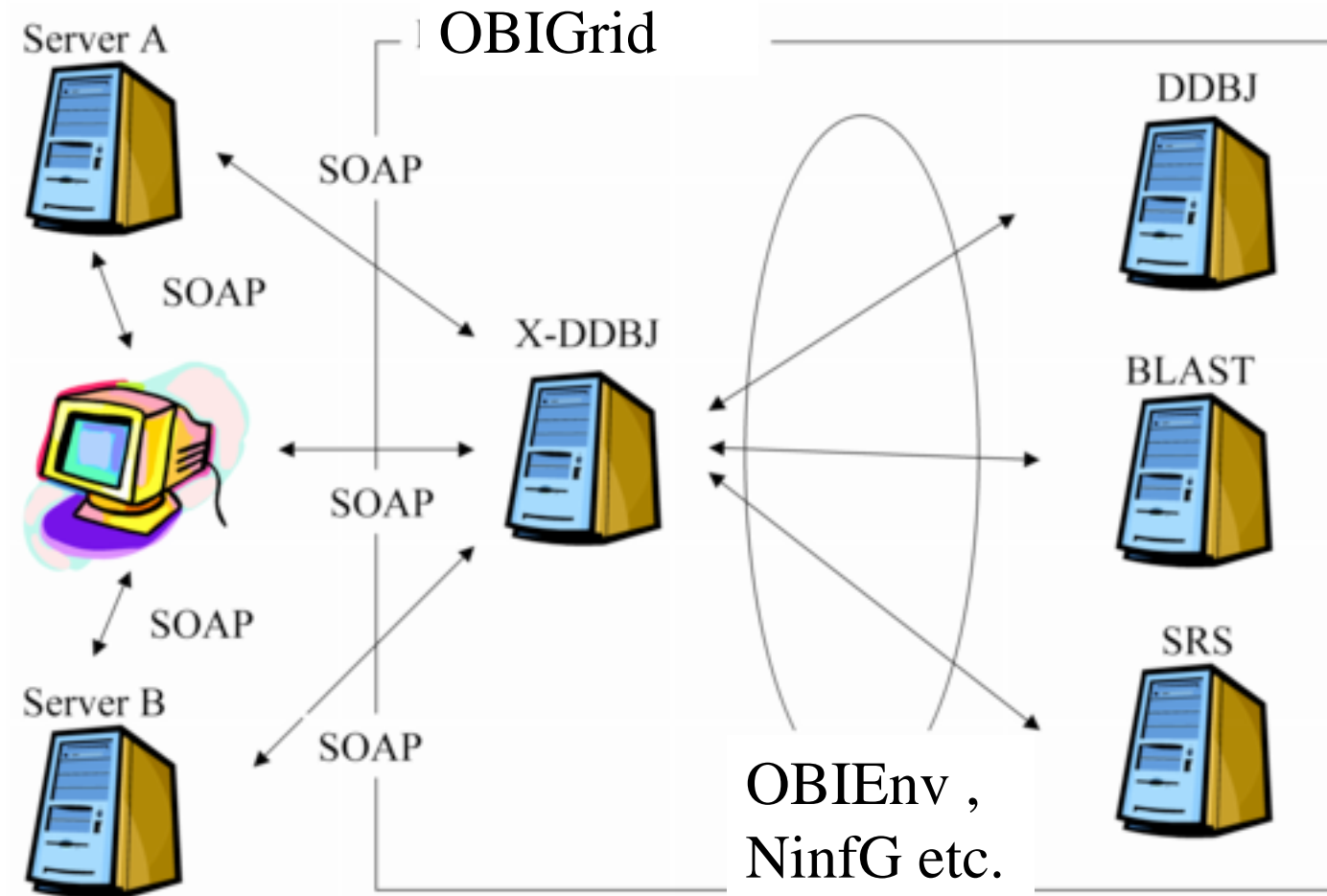
Homology search against multiple databases

Pattern matching against multiple databases

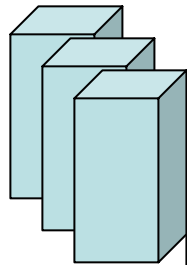
Multiple alignment/phylogenetic analysis

**Interactive and repetitive analysis and review
by annotators**

Test bed



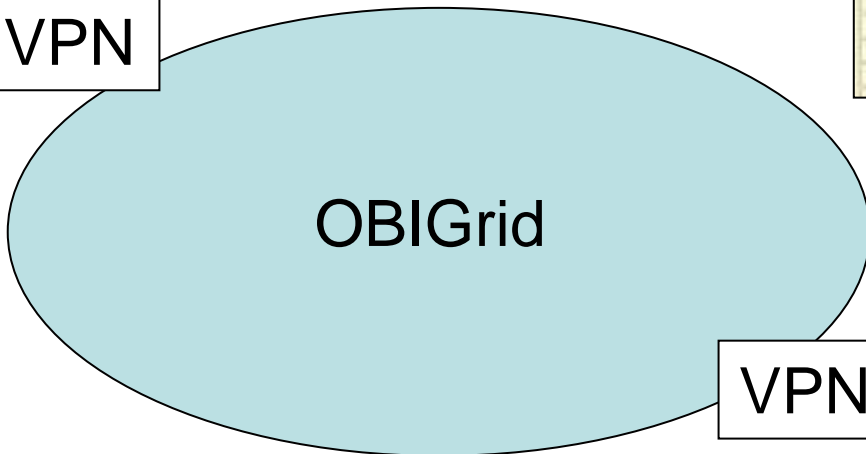
JAIST



Keyword search engine based on RDB

VPN

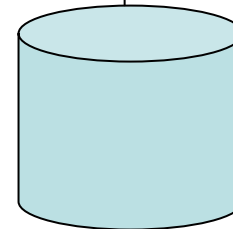
No.	Locus	Value 1	Value 2	Value 3	note
1	AF025413	lacZ			bc11
2	AF000141	lacZ			bc11
3	AF000385	regulator of lacZ			bc11
4	AF001102	lacZ expression regulatory protein (ec)			bc11
5	AF005213	lacZ			bc11
6	AF005533	regulator of lacZ			bc11
7	AF006428	lacZ			bc11
8	AF008846	similar to E. coli regulator of lacZ (AAC76068.1) [Blast hit to AAC76068.1 (275 aa) identity in aa 1-275]			bc11
9	AF011700	lacZ			bc12
10	AF013784	lacZ			bc12
11	AF013954	regulator of lacZ			bc12



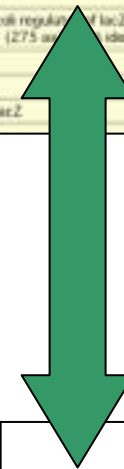
OBIGrid

VPN

X-DDBJ
SOAP server



NIG



Communication by SOAP

You are the sunshine of our project.

SOAP servers and Web services

YAMAGUCHI Masahito sun (Fujitsu Limited)

SHIGEMOTO Yasumasa sun (Fujitsu Limited)

MATSUO Masashi sun (Fujitsu Limited)

OBIGrid and OBI-Env linkage

KONAGAYA Akihiko sun (JAIST and RIKEN GSC)

SATOU Kenji sun (JAIST)

TSUJI Shinichi sun (JAIST)