Quality of services of a primary nucleotide sequence database

Hideaki Sugawara

Center for Information Biology and DDBJ,
National Institute of Genetics

SOKEN-DAI.

International Nucleotide Sequence Database Collaboration (INSDC) serves communities

JPO: Japan Patent Office

DDBJ: DNA Data Bank of Japan

NIG: National Institute of Genetics

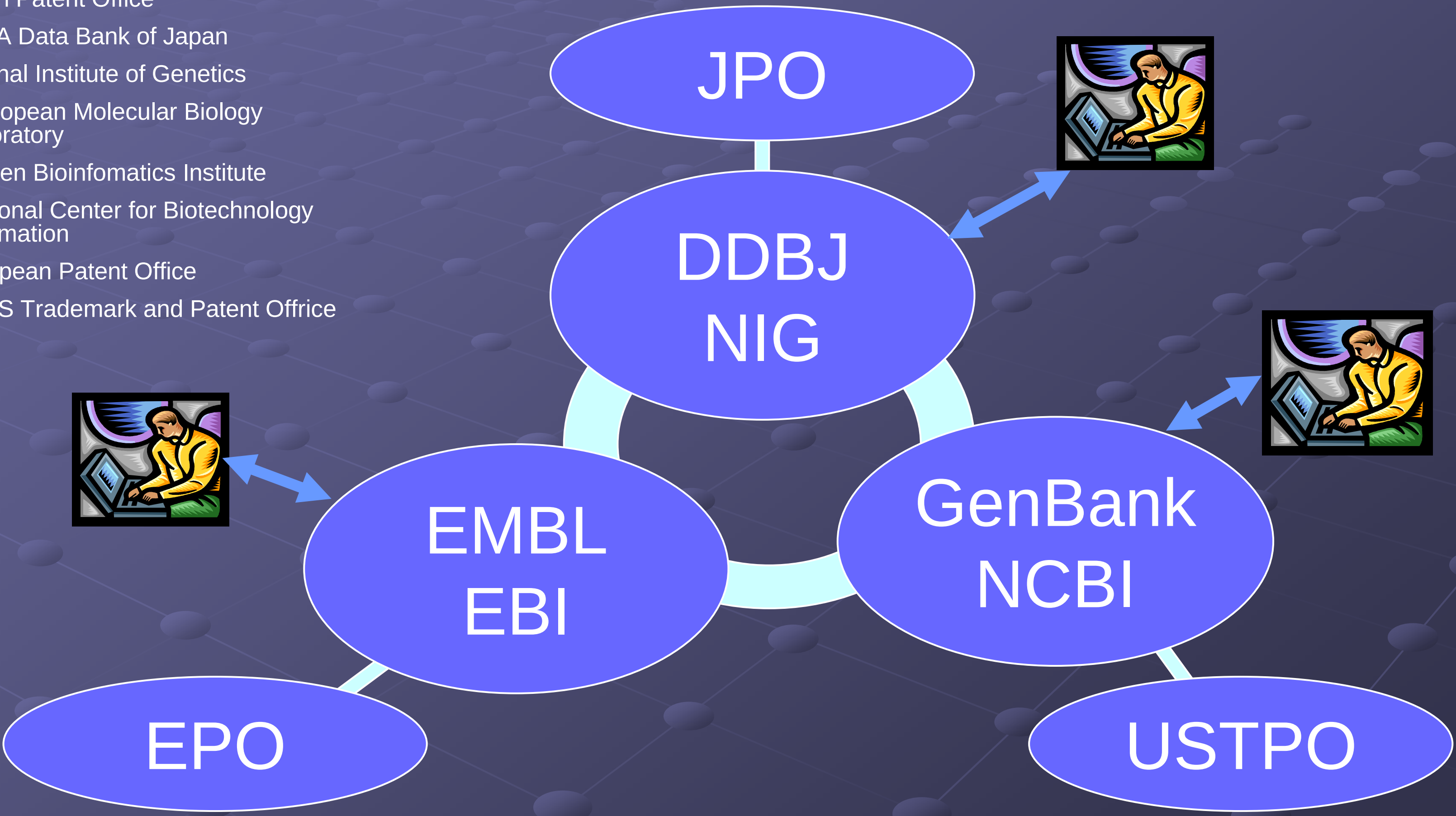
EMBL: European Molecular Biology Laboratory

EBI: Europen Bioinfomatics Institute

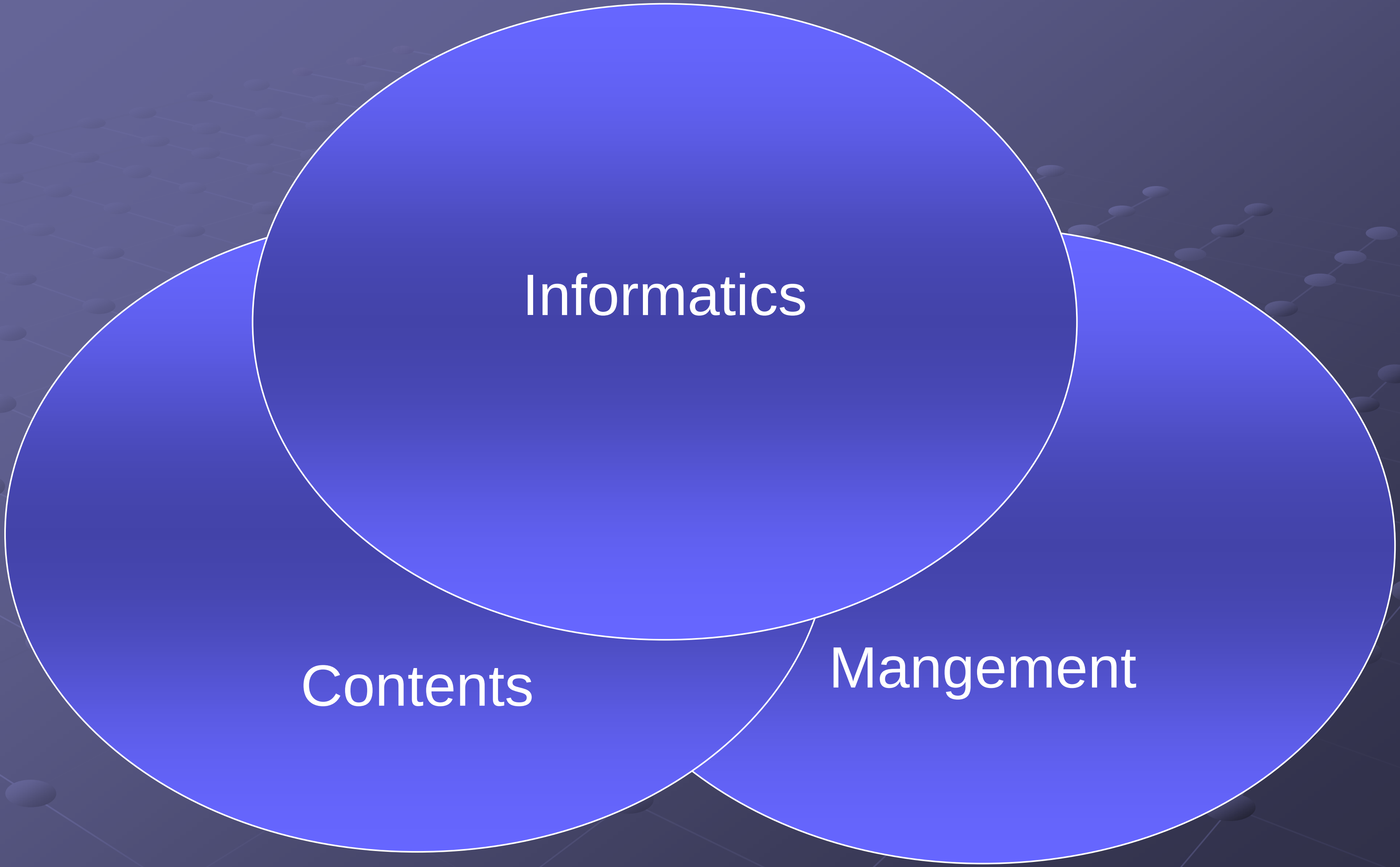
NCBI: National Center for Biotechnology Information

EPO: European Patent Office

USTPO: US Trademark and Patent Office



Infrastructure for the services

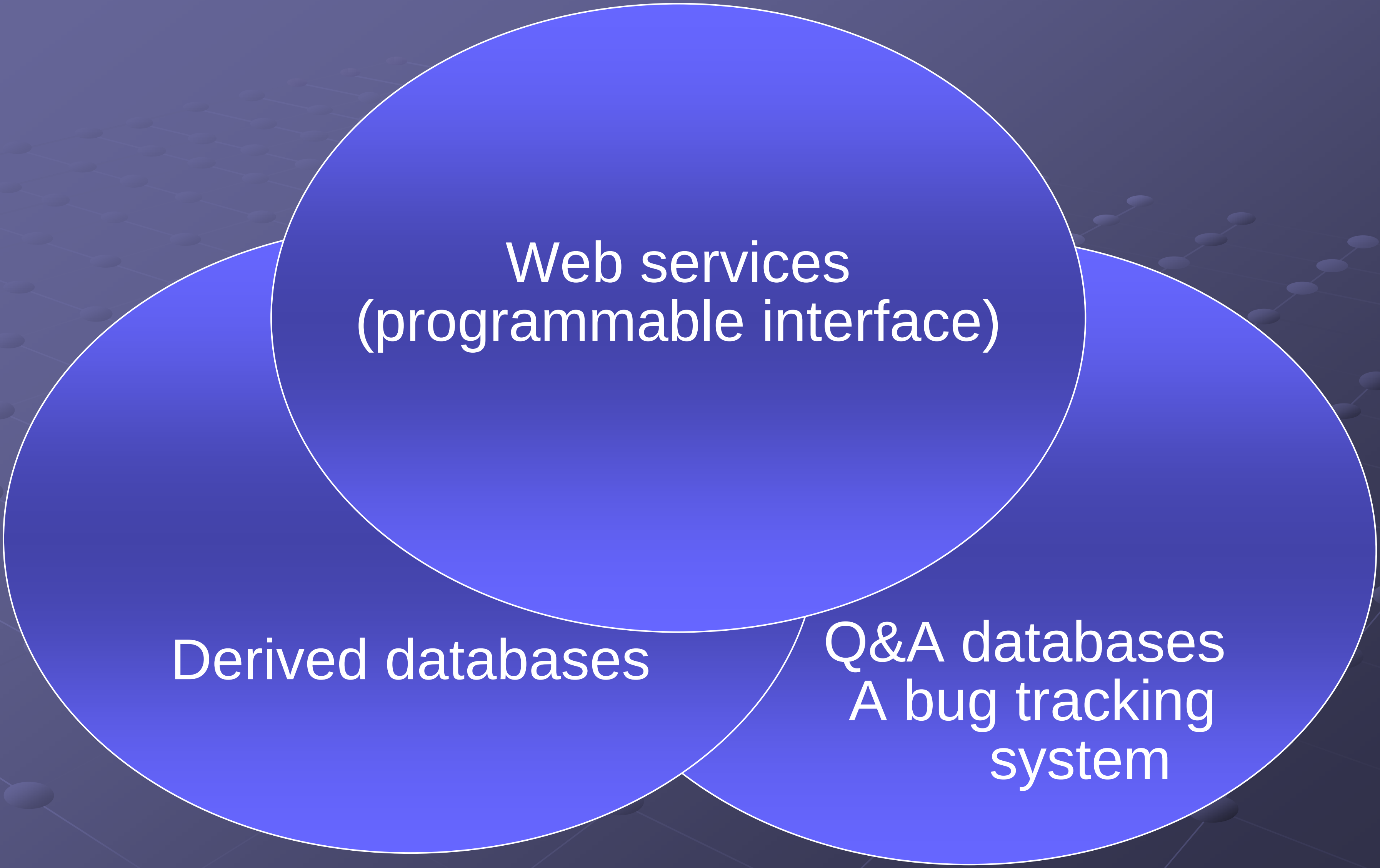


Informatics

Contents

Mangement

Today I will introduce:



Web services
(programmable interface)

Derived databases

Q&A databases
A bug tracking system

Mash-up of biological data resources

- **It has bee feasible to mash-up diverse databases and tools by hand to some extent.**

NCBI PubMed National Library of Medicine

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

- Enter one or more search terms.
- You may combine subject terms, author names and journal titles.

PubMed, a service of the National Library of Medicine, includes over 15 million citations from MEDLINE and additional life science journals for biomedical articles back to the 1950's. PubMed includes links to full text articles and other related resources.

Bookshelf PubMed Help

The PubMed Help Documentation is now available for interactive searching on the Bookshelf.

My NCBI Quick Tours

Learn how to save searches, get e-mail alerts and set up filters with the My NCBI Quick Tours.

A. Web site of Journal database

① Connect to a journal database to find accession numbers of INSDC

DNA Data Bank of Japan

English Page

Go

DDBJ とは

DDBJ は、DNA Data Bank of Japan の略称です。DDBJ は欧州の EBI/EMBL および米国の NCBI/GenBank との密接な連携のもと「DDBJ/EMBL/GenBank 国際塩基配列データベース」を構築している三大国際 DNA データバンクのひとつです。
静岡県三島市にある国立遺伝学研究所 生命情報・DDBJ 研究センター内で運営されています。

主な活動

- 国際塩基配列データベースの共同構築と運営
- 関連生命情報データベースの運営
- DNA データベースのオンライン利用の管理・運営
- ソフトウエアの開発
- 広報活動
- 国立遺伝学研究所コンピュータシステムならびにネットワークの管理・運用

B. Web site of Nucleotide database

②. Move to INSDC to search amino acid sequences by the accession numbers found in the step ①

PDBj
Protein Data Bank Japan

C. Web site of Protein Data Bank (PDB)

③. Move to the Protein Data Bank (PDB) to get the 3D structure.

Users with Web browser

- **Large-scale databases are produced in OMICS. Therefore, a machanical mash-up by program will be required more and more.**

Web services for the biological problem solving environment

We implemented a **SOAP (Simple Object Access Protocol) server** and **Web services** that provide a program-friendly interface.

We propose standardization of bioinformatics services to improve the interoperability of desperately diverse biological data resources.

The world of Web services

「Web services」 provider

EBI, NCBI, -----

DDBJ

DDBJ とは

DDBJ は，DNA Data Bank of Japan の略称です。DDBJ は欧州の EBI/EMBL および米国の NCBI/GenBank との密接な連携のもと「DDBJ/EMBL/GenBank 国際塩基配列データベース」を構築している三大国際 DNA データバンクのひとつです。
静岡県三島市にある国立遺伝学研究所 生命情報・DDBJ 研究センター内で運営されています。

主な活動

- 国際塩基配列データベースの共同構築と運営
- 関連生命情報データベースの運営
- DNA データベースのオンライン利用の管理・運営
- ソフトウエアの開発
- 広報活動
- 国立遺伝学研究所コンピュータシステムならびにネットワークの管理・運用

SOAP
(HTTP/HTTPS)

WSDL

XML

Registration of Web services

UDDI

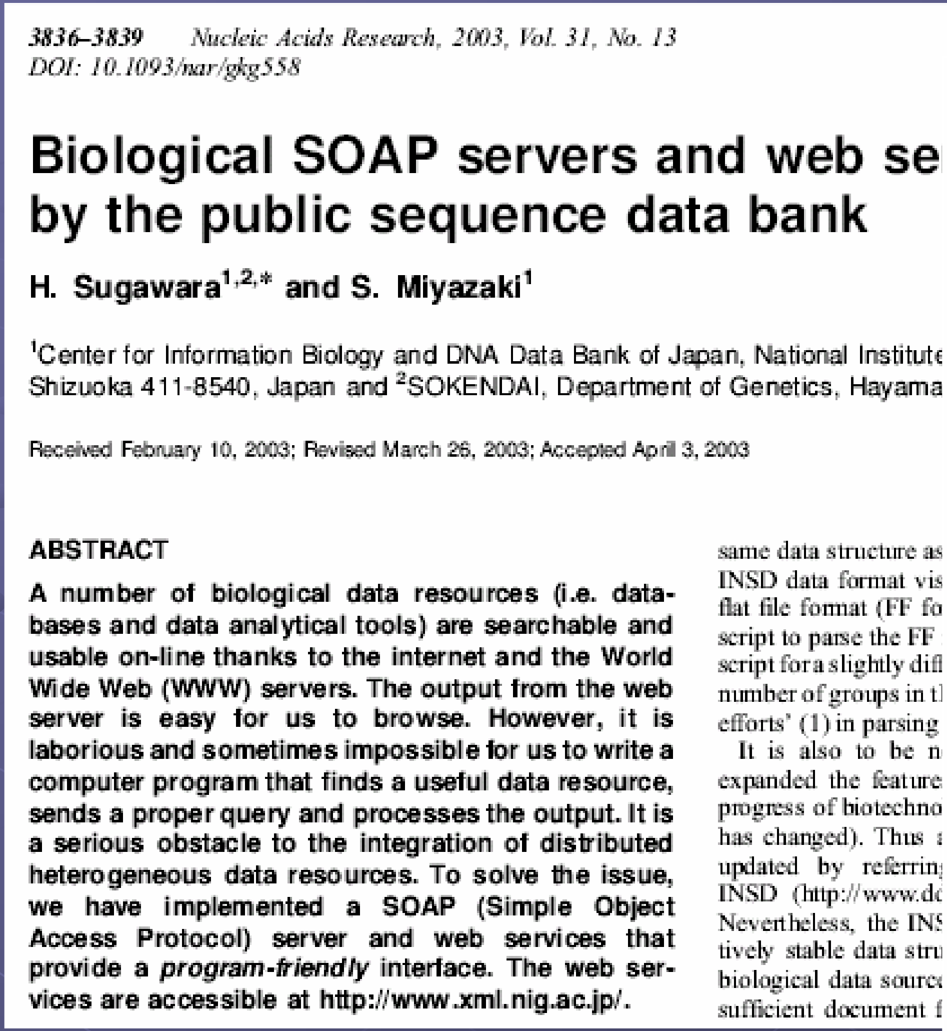
Search Web services

Users

WSDL
(Web Services Description Language).

UDDI (Universal Description, Discovery and Integration)

XML Central of DDBJ: http://xml.nig.ac.jp/index.html



3836–3839 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg558

Biological SOAP servers and web se
by the public sequence data bank

H. Sugawara[1,2,*] and S. Miyazaki[1]

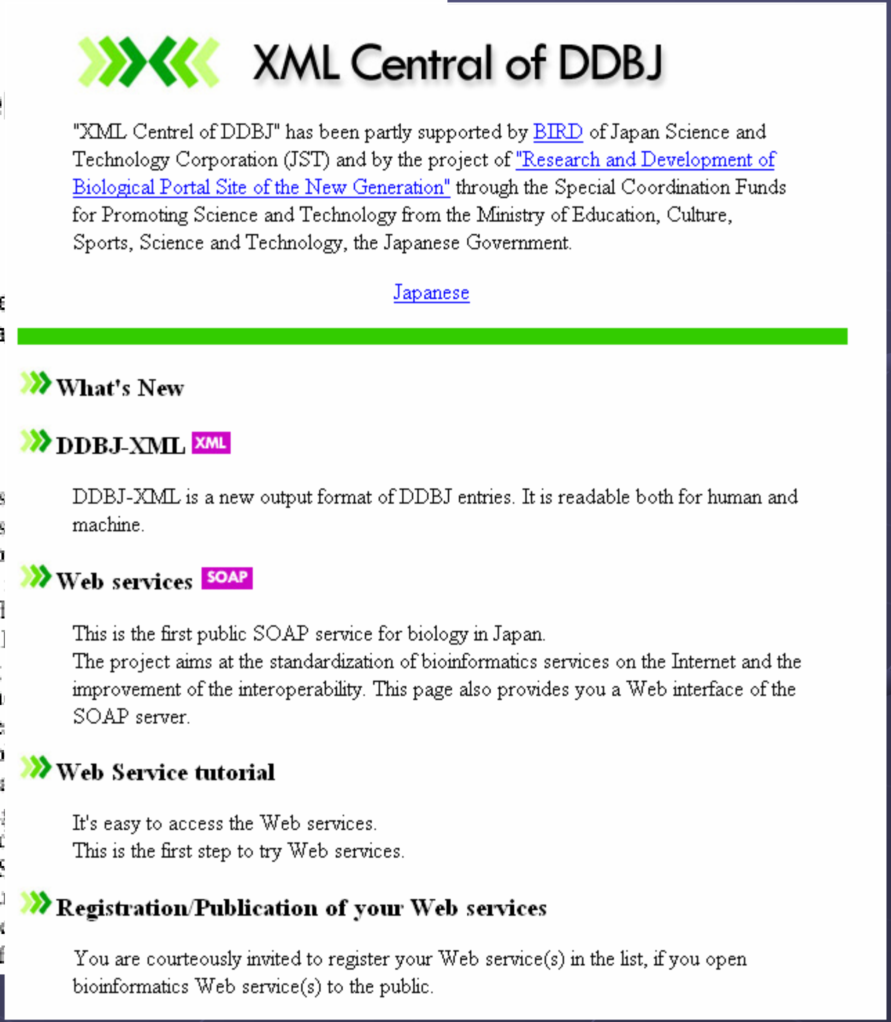
[1]Center for Information Biology and DNA Data Bank of Japan, National Institute
Shizuoka 411-8540, Japan and [2]SOKENDAI, Department of Genetics, Hayama

Received February 10, 2003; Revised March 26, 2003; Accepted April 3, 2003

ABSTRACT

A number of biological data resources (i.e. databases and data analytical tools) are searchable and usable on-line thanks to the internet and the World Wide Web (WWW) servers. The output from the web server is easy for us to browse. However, it is laborious and sometimes impossible for us to write a computer program that finds a useful data resource, sends a proper query and processes the output. It is a serious obstacle to the integration of distributed heterogeneous data resources. To solve the issue, we have implemented a SOAP (Simple Object Access Protocol) server and web services that provide a *program-friendly* interface. The web services are accessible at http://www.xml.nig.ac.jp/.

same data structure as
INSD data format vis
flat file format (FF fo
script to parse the FF
script for a slightly dif
number of groups in t
efforts' (1) in parsing
It is also to be n
expanded the feature
progress of biotechno
has changed). Thus
updated by referrin
INSD (http://www.d
Nevertheless, the INS
tively stable data stru
biological data source
sufficient document



XML Central of DDBJ

"XML Centrel of DDBJ" has been partly supported by BIRD of Japan Science and Technology Corporation (JST) and by the project of "Research and Development of Biological Portal Site of the New Generation" through the Special Coordination Funds for Promoting Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

Japanese

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.

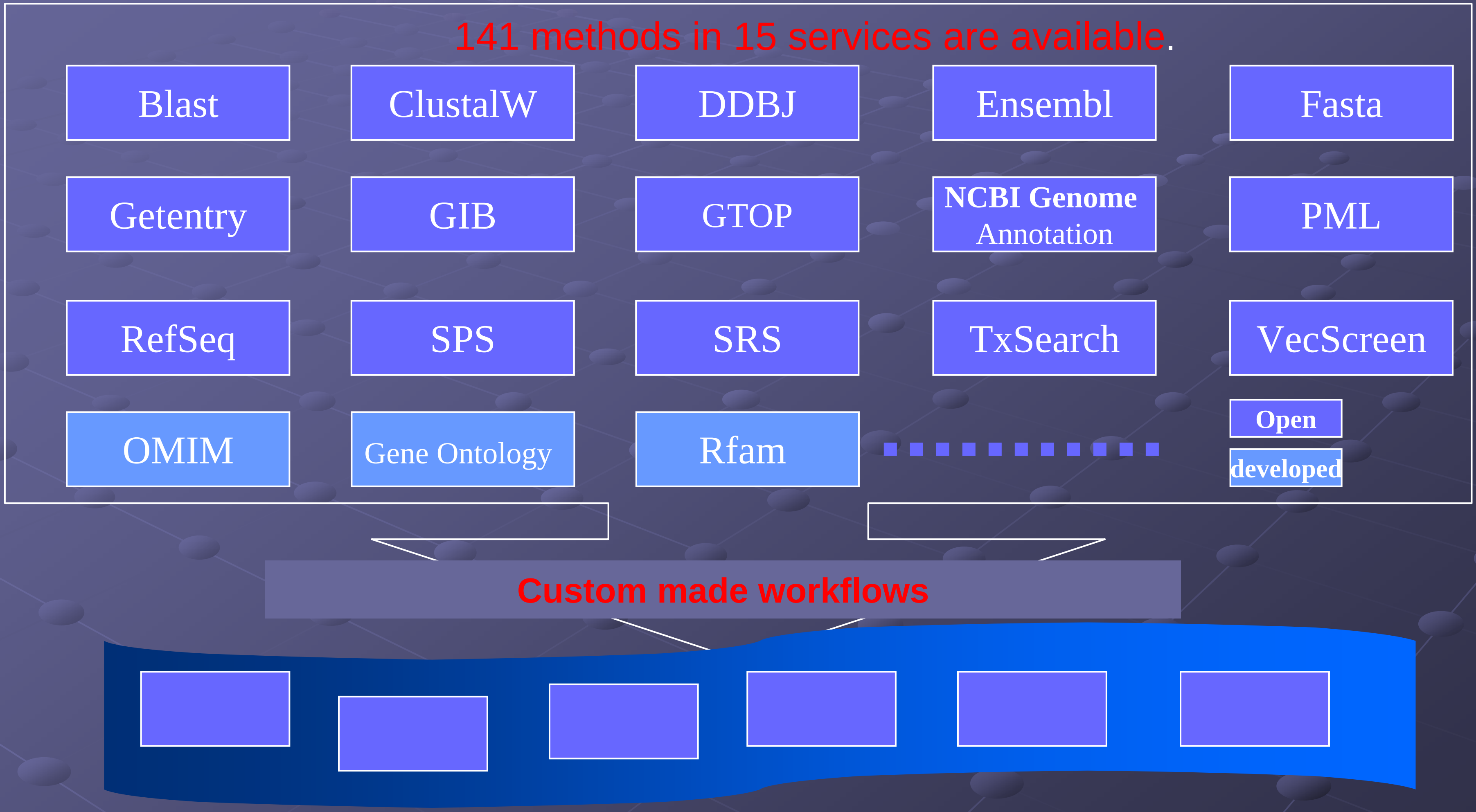
Web Service tutorial

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

Registration/Publication of your Web services

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

Workflow composed of methods provided by Web services



141 methods in 15 services are available.

Blast

ClustalW

DDBJ

Ensembl

Fasta

Getentry

GIB

GTOP

NCBI Genome Annotation

PML

RefSeq

SPS

SRS

TxSearch

VecScreen

OMIM

Gene Ontology

Rfam

Open

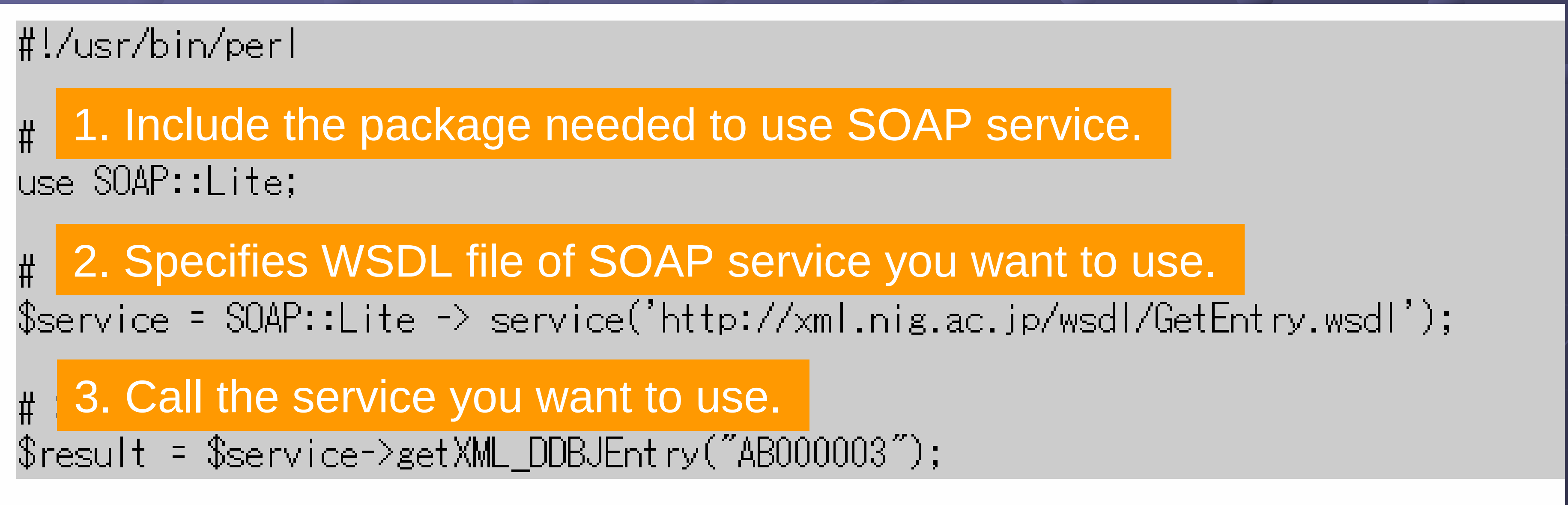
developed

Custom made workflows

Easy to bind Web services methods from your program

- **Download ActivePerl for WINDOWS and you can call methods in DDBJ Web services.**
- **Specify a WDSL file and call a method:**

  In the following example, the method of getXML_DDBJEntry(accession) of the Getentry Web services is called from a Perl program:



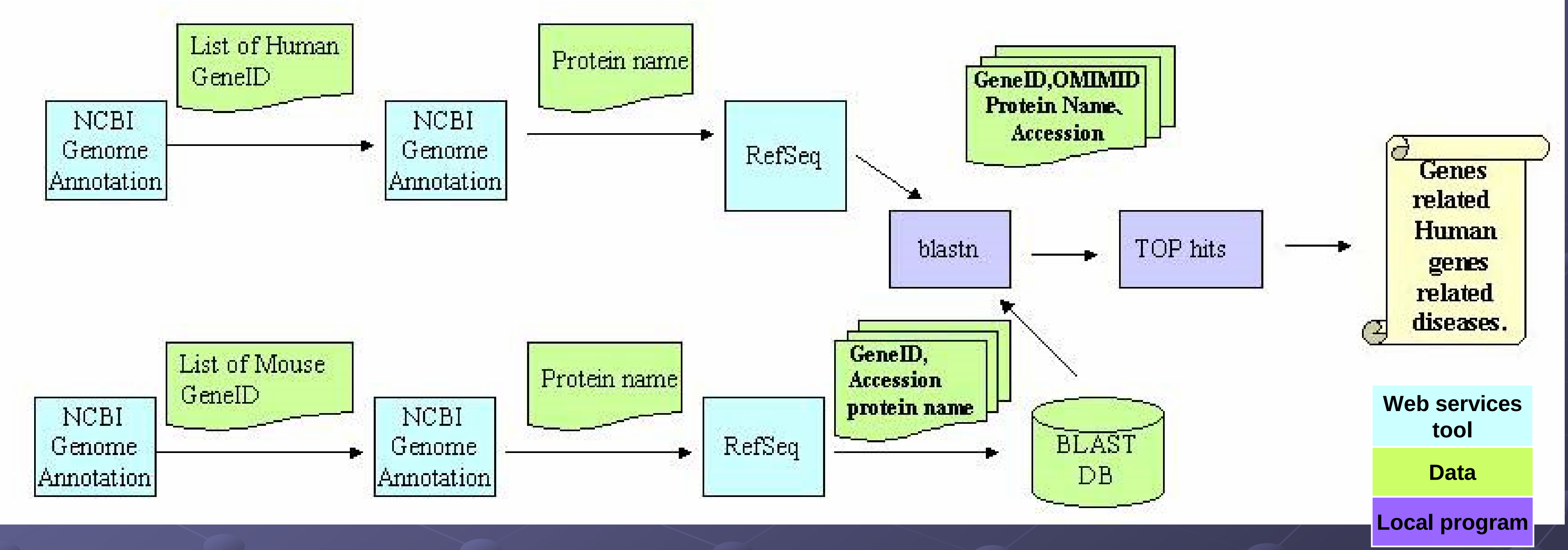
```
#!/usr/bin/perl

#
use SOAP::Lite;

#
$service = SOAP::Lite -> service('http://xml.nig.ac.jp/wsdl/GetEntry.wsdl');

#
$result = $service->getXML_DDBJEntry("AB000003");
```

1. Include the package needed to use SOAP service.

2. Specifies WSDL file of SOAP service you want to use.

3. Call the service you want to use.

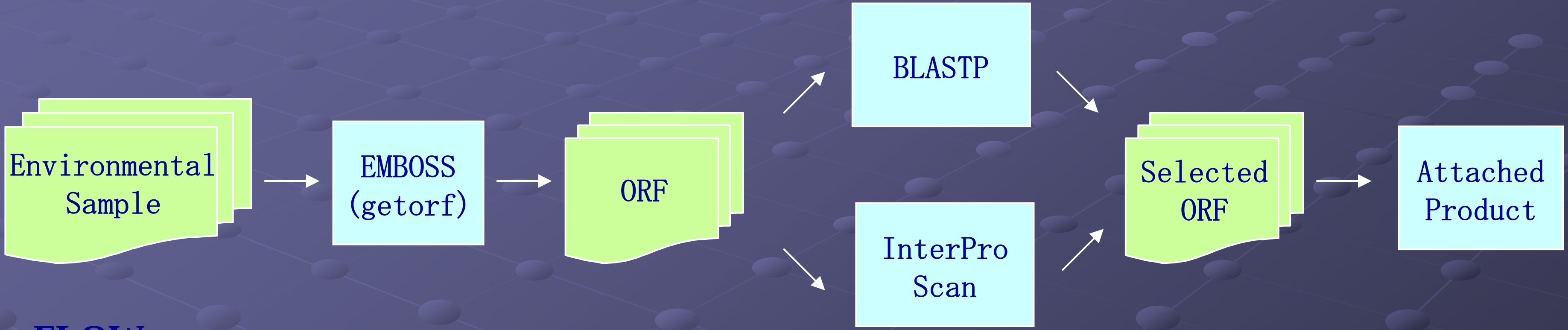
Workflow with Online Mendelian Inheritance in Man (OMIM)



This workflow reveal homology relationship between human disease genes and genes of other eukaryotes.

Environmental DNA sequences automatic annotation workflow

In order to make full use of gene information included in nucleotide sequence database, we developed workflow of gene finding of DNA fragments obtained from pooled genome samples of uncultured microbes in environmental samples.

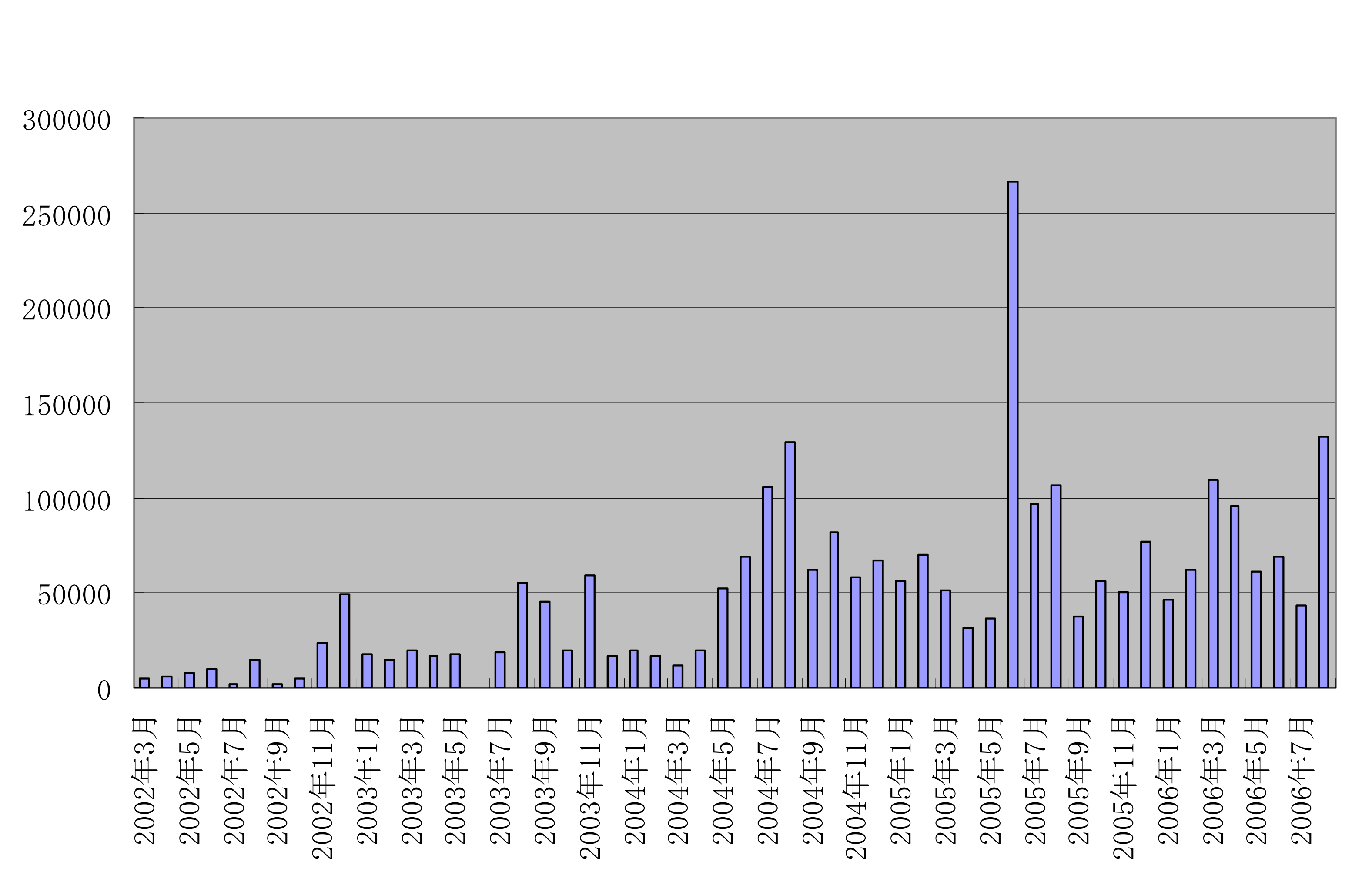


FLOW

1. Execute getorf in EMBOSS which finds and outputs the sequences of open reading frames (ORFs). The ORFs can be defined as regions of a specified minimum size (longer than 300bp) between START and STOP codons. (The minimum size cut-off for getorf was)
2. The protein product was attacheded by searching the bacterial division of DDBJ release 62 against predicted ORFs using BLASTP. (The e-value cut-off for BLASTP was 1e-40.)

Access frequency of DDBJ Web services

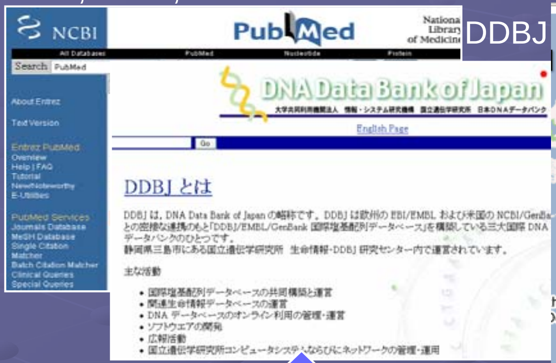
- Examples of methods frequently used: getentry, blast, RefSeq、GO



Web services in the world

EBI, NCBI, ----

DDBJ



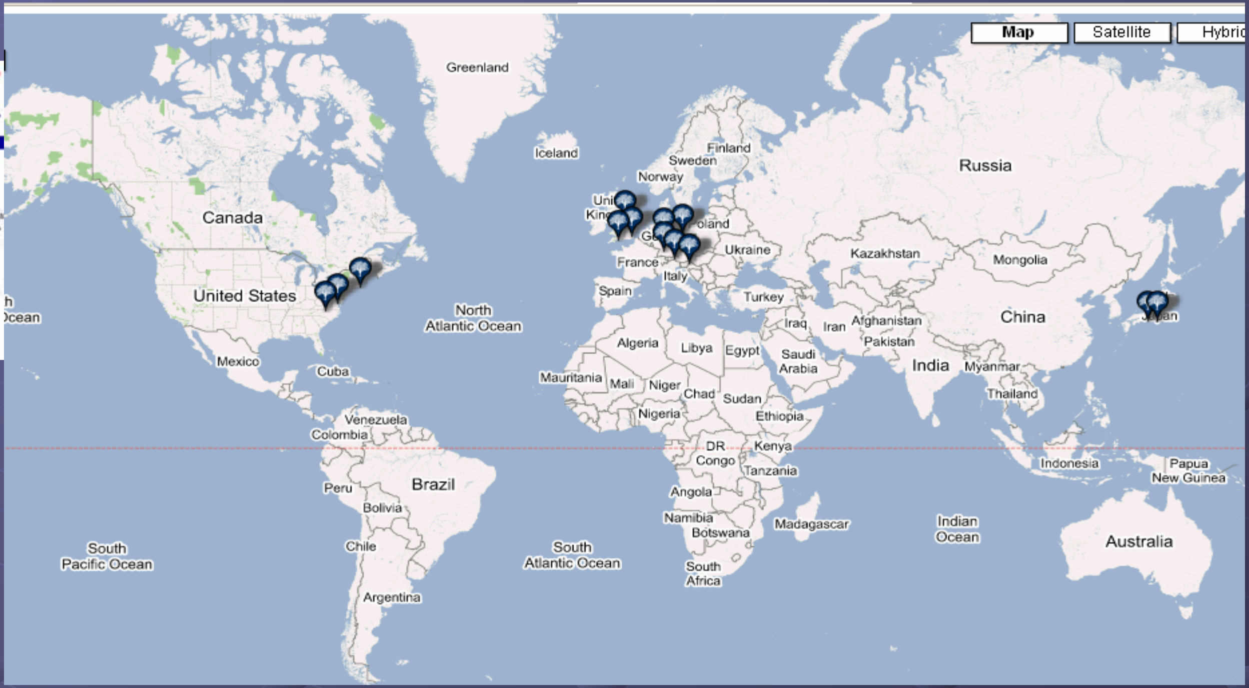
DDBJ とは

DDBJ は，DNA Data Bank of Japan の略称です。DDBJ は欧州の EBI/EMBL および米国の NCBI/GenBa
との密接な連携のもと「DDBJ/EMBL/GenBank 国際塩基配列データベース」を構築している三大国際 DNA データバンクのひとつです。
静岡県三島市にある国立遺伝学研究所　生命情報･DDBJ 研究センター内で運営されています。

主な活動

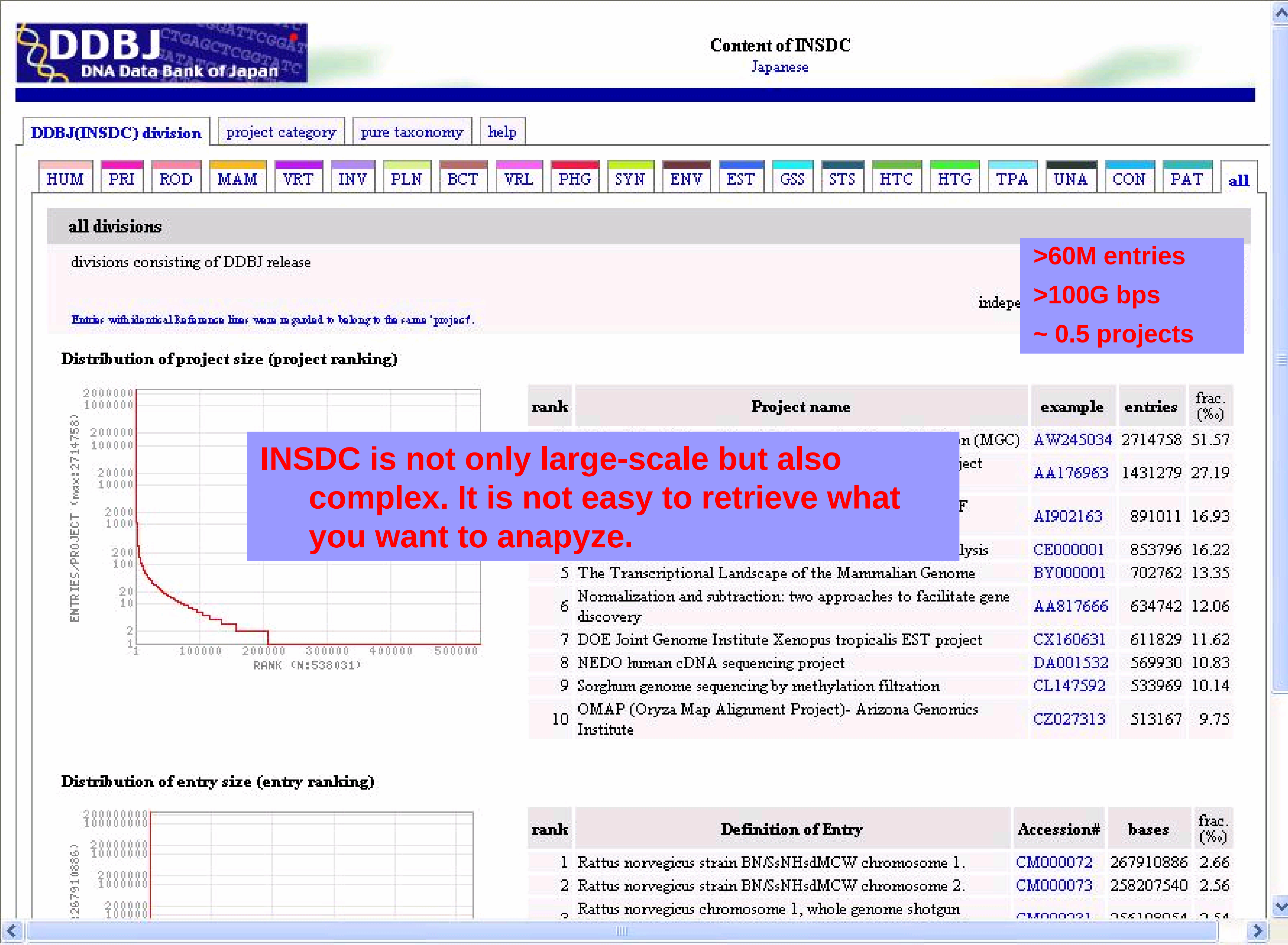
- 国際塩基配列データベースの共同構築と運営
- 関連生命情報データベースの運営
- DNA データベースのオンライン利用の管理･運営
- ソフトウエアの開発
- 広報活動
- 国立遺伝学研究所コンピュータシステムならびにネットワークの管理･運用

SOAP(HTTP/HTTPS)



(Web services map prepared by EBI)

Derived databases



Content of INSDC

all divisions

divisions consisting of DDBJ release

indepe

Entries with identical Reference lines were regarded to belong to the same 'project'.

>60M entries

>100G bps

~ 0.5 projects

Distribution of project size (project ranking)

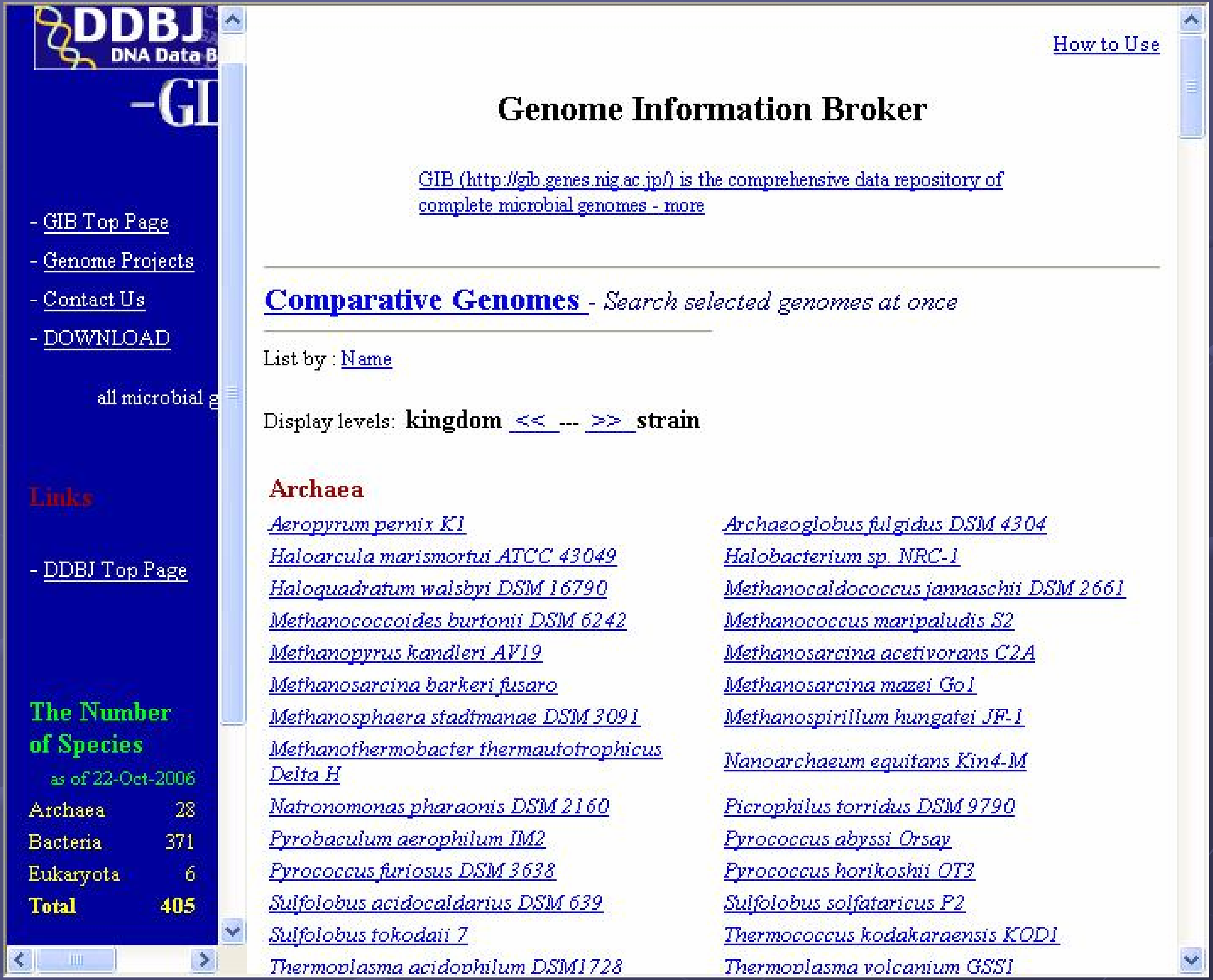
rank	Project name	example	entries	frac. (‰)
	n (MGC)	AW245034	2714758	51.57
	ject	AA176963	1431279	27.19
	F	AI902163	891011	16.93
	lysis	CE000001	853796	16.22
5	The Transcriptional Landscape of the Mammalian Genome	BY000001	702762	13.35
6	Normalization and subtraction: two approaches to facilitate gene discovery	AA817666	634742	12.06
7	DOE Joint Genome Institute Xenopus tropicalis EST project	CX160631	611829	11.62
8	NEDO human cDNA sequencing project	DA001532	569930	10.83
9	Sorghum genome sequencing by methylation filtration	CL147592	533969	10.14
10	OMAP (Oryza Map Alignment Project)- Arizona Genomics Institute	CZ027313	513167	9.75

INSDC is not only large-scale but also complex. It is not easy to retrieve what you want to anapyze.

Distribution of entry size (entry ranking)

rank	Definition of Entry	Accession#	bases	frac. (‰)
1	Rattus norvegicus strain BN/SsNHsdMCW chromosome 1.	CM000072	267910886	2.66
2	Rattus norvegicus strain BN/SsNHsdMCW chromosome 2.	CM000073	258207540	2.56
3	Rattus norvegicus chromosome 1, whole genome shotgun	CM000231	256109054	2.54

A simple derived database: subset of complete microorganisms genomes from INSDC



DDBJ
DNA Data B
-GI

- GIB Top Page
- Genome Projects
- Contact Us
- DOWNLOAD

all microbial g

Links

- DDBJ Top Page

The Number of Species
as of 22-Oct-2006

Archaea	28
Bacteria	371
Eukaryota	6
Total	**405**

How to Use

Genome Information Broker

GIB (http://gib.genes.nig.ac.jp/) is the comprehensive data repository of complete microbial genomes - more

Comparative Genomes - *Search selected genomes at once*

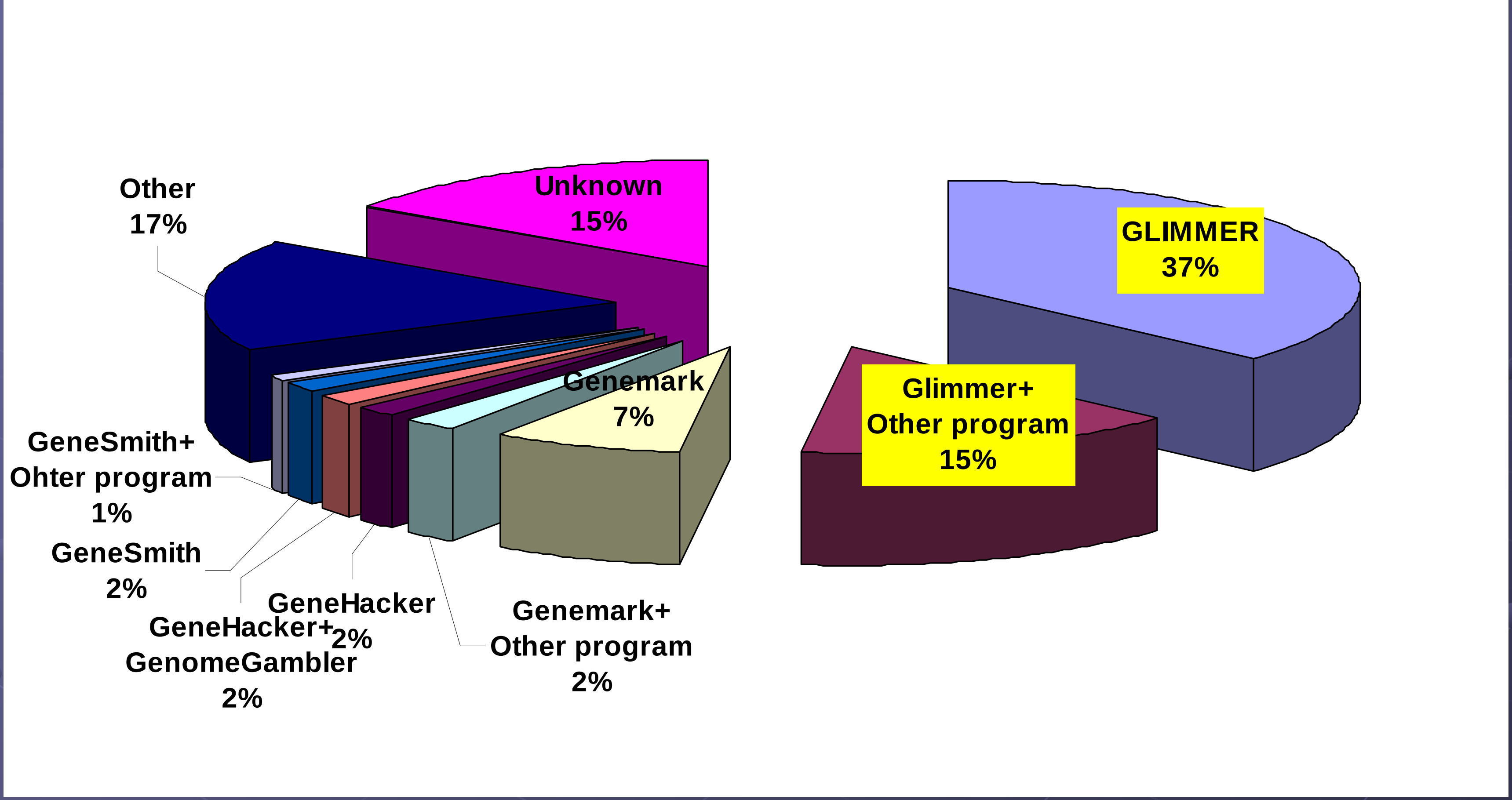
List by : Name

Display levels: **kingdom** << --- >> **strain**

Archaea

Aeropyrum pernix K1
Archaeoglobus fulgidus DSM 4304
Haloarcula marismortui ATCC 43049
Halobacterium sp. NRC-1
Haloquadratum walsbyi DSM 16790
Methanocaldococcus jannaschii DSM 2661
Methanococcoides burtonii DSM 6242
Methanococcus maripaludis S2
Methanopyrus kandleri AV19
Methanosarcina acetivorans C2A
Methanosarcina barkeri fusaro
Methanosarcina mazei Go1
Methanosphaera stadtmanae DSM 3091
Methanospirillum hungatei JF-1
Methanothermobacter thermautotrophicus Delta H
Nanoarchaeum equitans Kin4-M
Natronomonas pharaonis DSM 2160
Picrophilus torridus DSM 9790
Pyrobaculum aerophilum IM2
Pyrococcus abyssi Orsay
Pyrococcus furiosus DSM 3638
Pyrococcus horikoshii OT3
Sulfolobus acidocaldarius DSM 639
Sulfolobus solfataricus P2
Sulfolobus tokodaii 7
Thermococcus kodakaraensis KOD1
Thermoplasma acidophilum DSM1728
Thermoplasma volcanium GSS1

Gene prediction programs used



Parameters: the cutoff length used

length	number
＞20	1
＞ （＝） 30	25
>33.3aa （100bp)	3
>40aa	1
>50aa	7
>60aa	4
>66.6aa (200bp)	1
>80	2
>100aa	6
>150aa	1
>200aa	1
>300aa	1
>400aa	1

Annotation: description of products

~ *Hahella chejuensis* KCTC 2396

```
     CDS             1023521..1024429
                     /gene="argB"
                     /locus_tag="HCH_01027"
```

COG ID in the

~ *Archaeoglobus fulgidus* DSM 4304

```
 CDS              complement(1141715..1142587)
                      /locus_tag="AF_1280"
                      /note="similar to GB:L77117 SP:Q60382 PID:1592260 percent
                      identity: 56.06; identified by sequence similarity;
```

Some details of
h
ana
/no

~ *Agrobacterium tumefaciens* C58 circular chromosome

```
     CDS              complement(373582..374466)
                      /gene="AGR_C_666"
                      /note="acetylglutamate kinase PA5323 {imported} -
                      Pseudomonas aeruginosa (strain PA01)"
                      /codon_start=1
                      /transl_table=11
                      /product="AGR_C_666p"
                      /protein_id="AAK86197.1"
                      /db_xref="GI:15155294"
                      translation="MTSSESEIQARLLAQALPFMQKYENKTIVVKYGGHAMGDSTLGK
                      FAEDIALLKQSGINPIVVHGGGPQIGAMLSKMGIESKFEGGLRVTDAKTVEIVEMVL
                      GSINKEIVALINQTGEWAIGLCGKDGNMVFAEKAKKTVIDPDSNIERVLDLGFVGEV
                      EVDRTLLDLLAKSEMIPVIAPVAPGRDGATYNINADTFAGAIAGALHATRLLFLTDV
                      PGVLDKNKELIKELTVSEARALIKDGTISGGMIPKVETCIDAIKAGVQGVVILNGKTP
                      HSVLLEIFTEGAGTLIVP"
```

Product name in the qualifier of /note

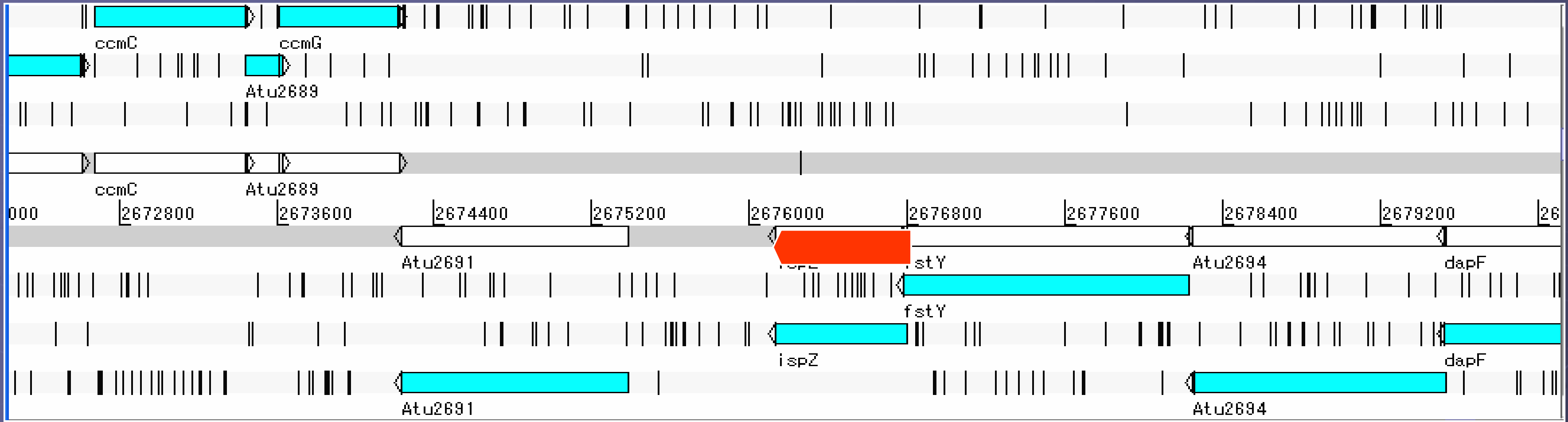
Product ID in the qualifier of /product

Annotation: description of products

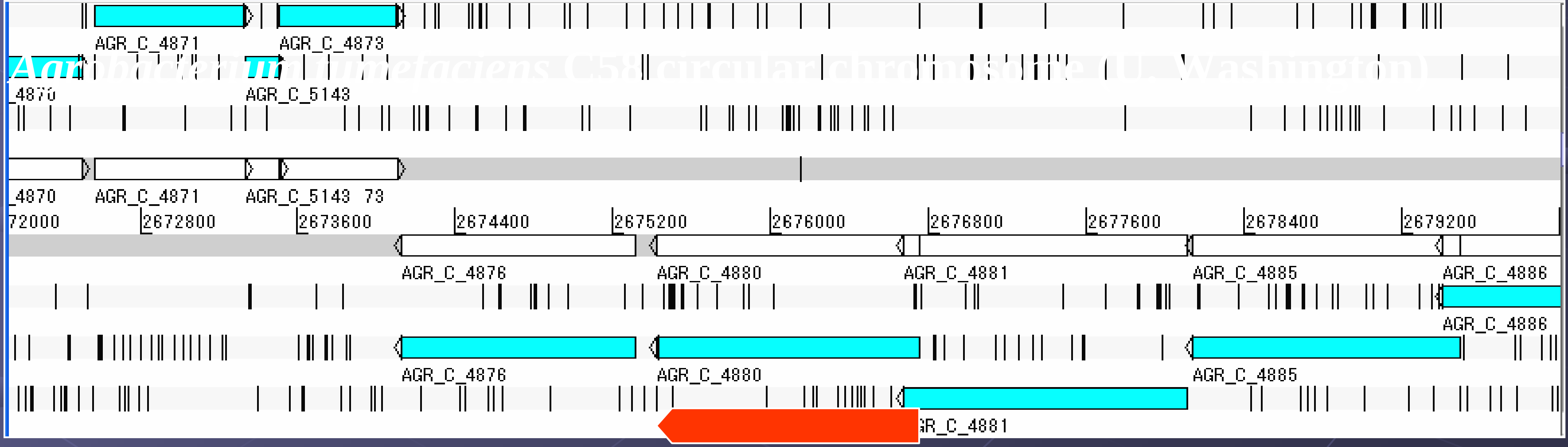
- unknown
- hypothetical protein
- probable orf
- predicted protein
- putative protein
- hypothetical conserved protein
- uncharacterized protein
- conserved domain protein
- conseved hypthetical

Annotation: inconsistency or biological variation

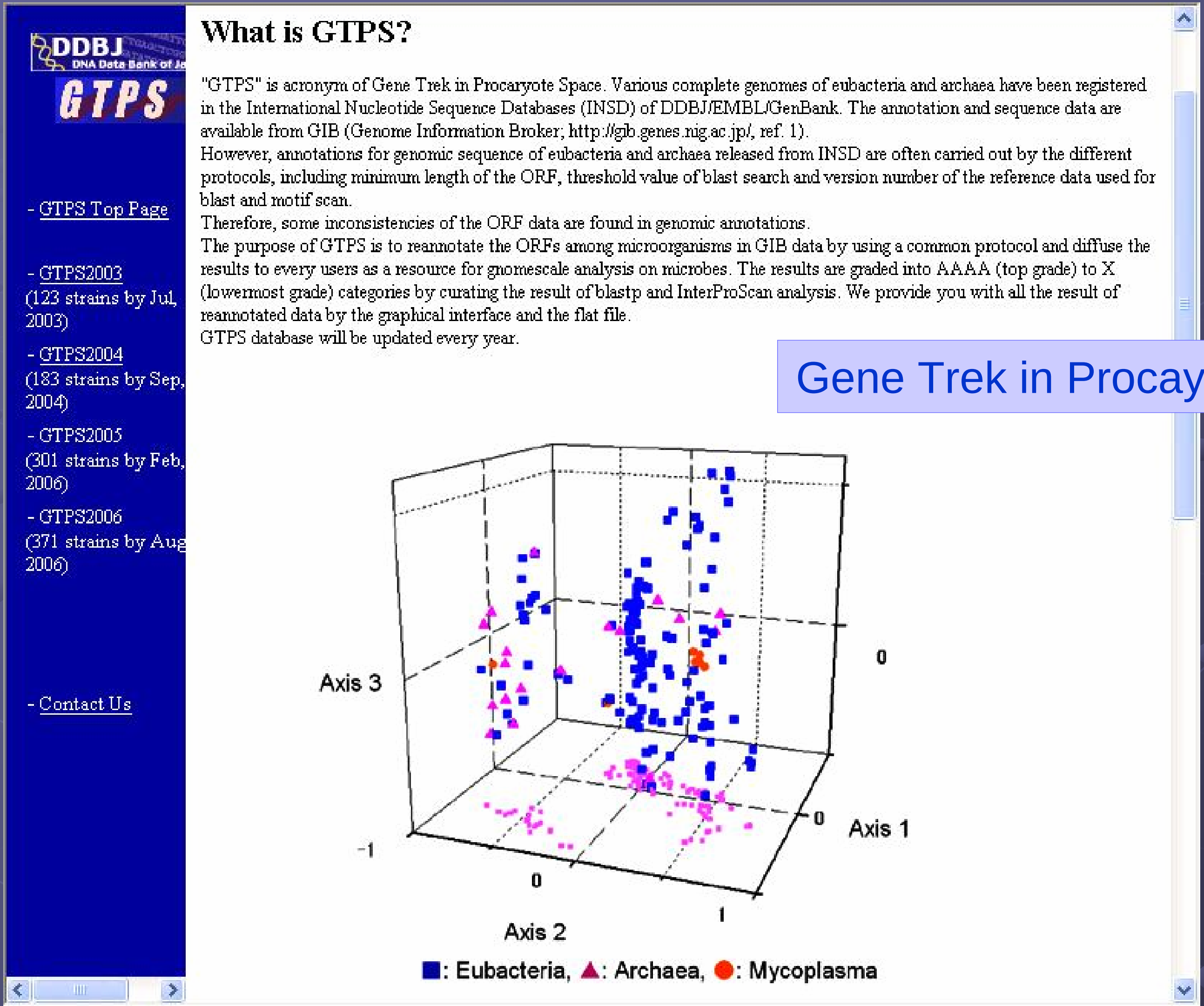
Agrobacterium tumefaciens C58 circular chromosome (Cereon)



Agrobacterium tumefaciens C58 circular chromosome (U. Washington)



A derived database: contents of complete microorganisms genomes were evaluated


DDBJ
DNA Data Bank of Ja
GTPS
- GTPS Top Page
- GTPS2003
(123 strains by Jul, 2003)
- GTPS2004
(183 strains by Sep, 2004)
- GTPS2005
(301 strains by Feb, 2006)
- GTPS2006
(371 strains by Aug 2006)
- Contact Us

What is GTPS?

"GTPS" is acronym of Gene Trek in Procaryote Space. Various complete genomes of eubacteria and archaea have been registered in the International Nucleotide Sequence Databases (INSD) of DDBJ/EMBL/GenBank. The annotation and sequence data are available from GIB (Genome Information Broker; http://gib.genes.nig.ac.jp/, ref. 1).
However, annotations for genomic sequence of eubacteria and archaea released from INSD are often carried out by the different protocols, including minimum length of the ORF, threshold value of blast search and version number of the reference data used for blast and motif scan.
Therefore, some inconsistencies of the ORF data are found in genomic annotations.
The purpose of GTPS is to reannotate the ORFs among microorganisms in GIB data by using a common protocol and diffuse the results to every users as a resource for gnomescale analysis on microbes. The results are graded into AAAA (top grade) to X (lowermost grade) categories by curating the result of blastp and InterProScan analysis. We provide you with all the result of reannotated data by the graphical interface and the flat file.
GTPS database will be updated every year.

Gene Trek in Procayote Space

Features of GTPS

- Simultaneous prediction and evaluation of all the possible protein coding genes (ORFs) in prokaryote genomes
- All the ORFs are graded
  - The criteria and evidence data are available
- Web site
  - http://gtps.ddbj.nig.ac.jp/
- Updated once a year

Short history of GTPS

ver. (data froze in)	strains	Archaea	Bacteria
2003 (Jul 2003)	123	14	109
2004 (Sep 2004)	183	17	166
2005 (Feb 2006)	302	25	277

Potential genes
AAAA1-D3 grades

Grade		blastp hit		InterProScan hit
		Coverage	Subject	Subject
AAAA	BBBB	alignment subject ⇔query ≧ 70% & or alignment/ORF ≧ 70%	Not putative membrane nor unknown protein	Function known motif
AAA	BBB			Unknown motif
AA	BB			No hit
A	B		Putative membrane or unknown protein	Function known or unknown motif
C		≧ 70% or	Putative membrane protein	No hit
		No hit		Function known or unknown motif
D		≧ 70% &	Unknown protein	No hit
E		≧ 70% or	Unknown protein	No hit
X		No hit		No hit

Number of ORFs sorted by each grade

Grades	GTPS	GTPS	GTPS
	ver. 2003	ver. 2004	ver. 2005
AAAA-A	283,247	431,672	752,186
BBBB-B	7,208	10,250	20,755
C	4,680	7,511	16,227
D	79,779	107,382	137,656
E	6,788	10,225	17,739
X	466,681	687,110	278,075
Total	848,383	1,254,150	1,222,638
	Glimmer2 and RBS finder		Glimmer 3
Potential genes (AAAA1 - D3)	370,876	551,246	903,845
INSDC	362,828	537,312	904,530

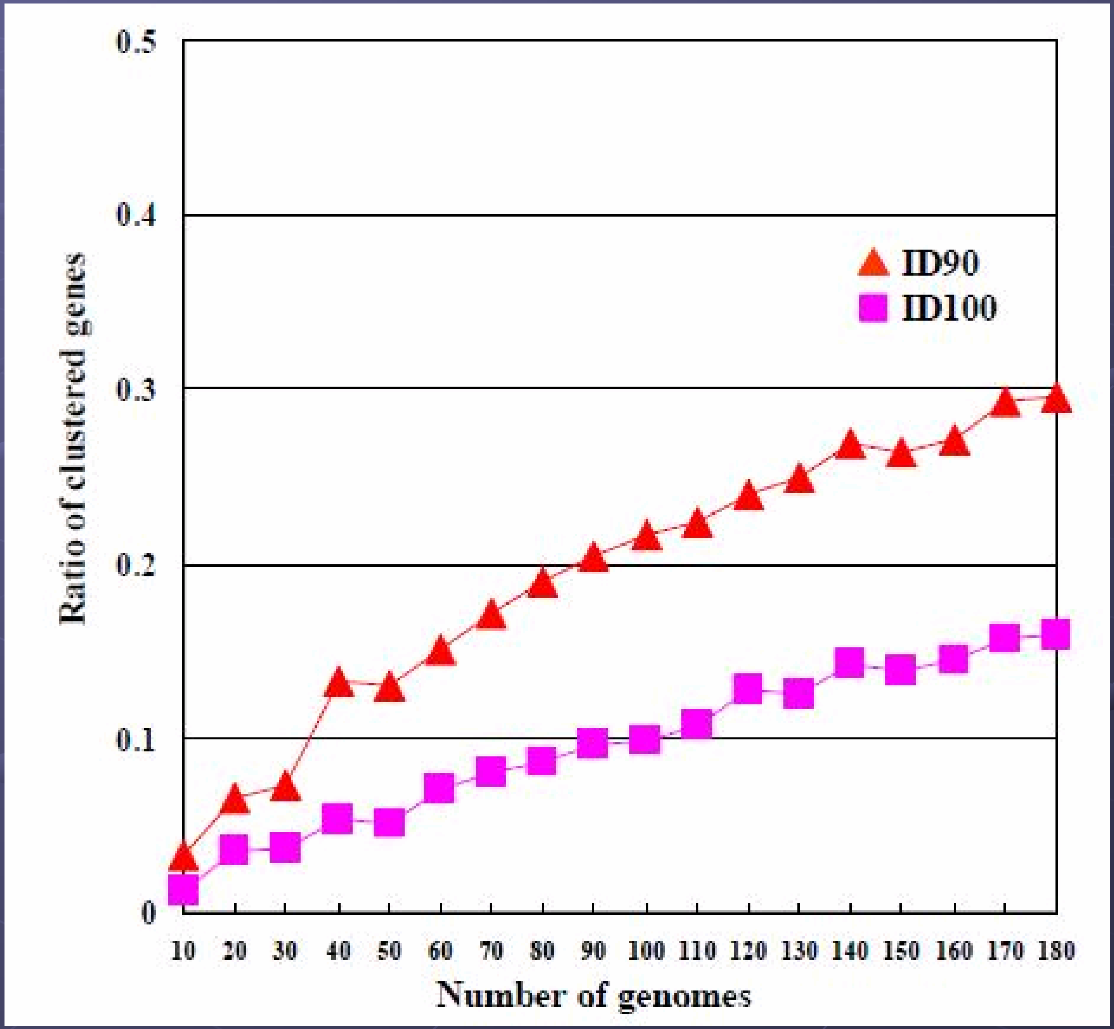
Comparison of the number of protein coding genes between GTPS and INSDC

	GTPS		GTPS		GTPS	
	ver. 2003		ver. 2004		ver. 2005	
Identical	261,720	70.6%	390,557	70.8%	576,762	63.8%
3' matched	92,206	24.9%	133,146	24.2%	252,365	27.9%
New ORFs (Not annotated in INSDC)	14,954	4.0%	23,935	4.3%	31,438	3.5%
Not predicted by Glimmer	1,996	0.5%	3,608	0.7%	43,280	4.8%
Total (potential genes)	370,876		551,246		903,845	

Glimmer 3 was used for ver.2005.

How many genes are out there?

The ratio of the clustered ORFs to all the ORFs among the sampled genomes increased with the increasing number of the genomes. The ratio is not yet saturated at the 180 genomes (29.5% of the genes among 180 genomes were clustered and presumed to have the same function.).



Summary

- It is the long term mission of the primary database to archive all the data published for the long term.
- At the same time, it is requested to provide *objective/reliable* data and tools for the *mash-up*.
- It proposes and maintains standards of biological data processing for the long term.