F8: DDBJ Activities: Contribution to the Research in Information Biology

Jun Mashima, Takehide Kosuge, Toshihisa Okido, Yuichi Kodama, Kazuho Ikeo, Hideaki Sugawara, Yoshio Tateno and Takashi Gojobori
Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Japan

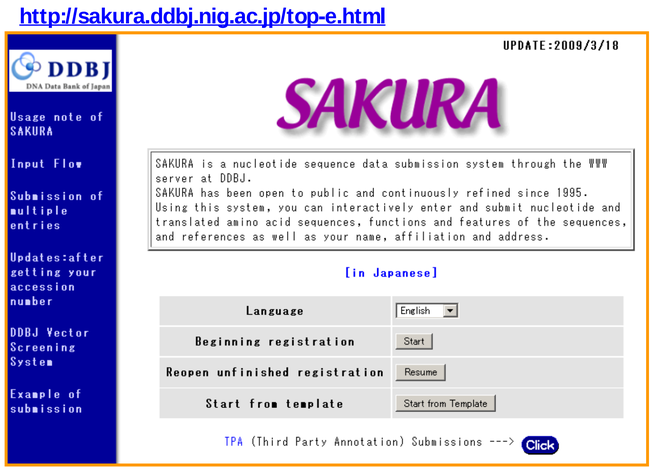
DDBJ
DNA Data Bank of Japan

Nucleotide Sequence Data
INSDC (DDBJ/EMBL-Bank/GenBank)

SAKURA

Interactive application for nucleotide sequence submission to enter all of items required for submission on a step-by-step basis

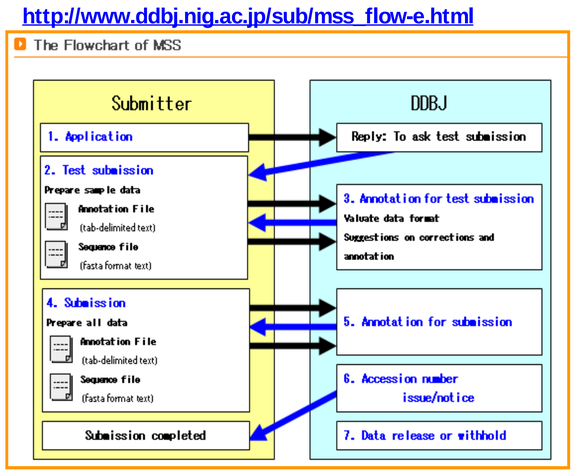
http://sakura.ddbj.nig.ac.jp/top-e.html



MSS (Mass Submission System)

- Large number of sequences (e.g. EST data)
- Complex submission with many features
- Long sequences (e.g. genome)
- Data for WGS, MGA etc.

http://www.ddbj.nig.ac.jp/sub/mss_flow-e.html



DDBJ is a member of INSDC. While many biological databases are constructed on data derived from the primary literature, most of sequence data submitted to INSDC are directly submitted from general researchers as a prerequisite to publication (required by most journals).

In direct submissions from general biologists there are many errors not only formative or careless mistakes but also scientific or logical misunderstandings. So, to find them manual curation is one of the most important jobs for DDBJ. In principle, DDBJ annotators review all sequence data submitted to DDBJ to suggest error corrections for submitters.

History of Annotation Jamborees
DDBJ Annotators Participated/Coorganized

FANTOM: Mouse cDNA	2000
FANTOM2: Mouse cDNA	2002
H-Inv1: Human cDNA	2002
E. coli Bioinformatics Meeting	2003
H-Inv2: Human cDNA	2003
E. coli annotation workshop	2003
RAP1: Rice genome and cDNA	2004
E. coli annotation workshop II	2005
RAP2: Rice genome and cDNA	2006
Theileria genome annotation	2007-2008

DDBJ annotators sometimes collaborate with large scale annotation projects.
H-Invitational and H-Invitational 2 were coorganized by JBIRC and DDBJ.

Nucleotide Number Collected by INSDC

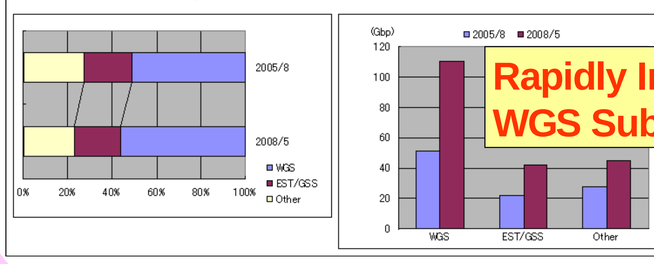
1986～2005 100 Gbp over
1986～2008 200 Gbp over

x 2, only in 3 years!

In May 2008, the total number of bases (DNA and RNA) collected and distributed by INSD (International Nucleotide Sequence Database: DDBJ/EMBL/GenBank) has reached 200 G bases (200,000,000,000 bases; the 'letters' of the genetic code). It took only three years from when we had reached 100G bases on August 2005.

INSD has expanded its specifications to accept data submissions from large scale sequencing projects. For example, we have started accepting the sequence data from EST (Expressed Sequence Tags) projects into EST division since 1993. On 2002, to accept submissions of the draft genome and the meta-genome sequences, we have created the new category for WGS (Whole Genome Shotgun) data.

The left figure shows the relationship of the ratios of three categories. The right figure shows the numbers of bases.



Rapidly Increasing WGS Submissions!

http://www.ddbj.nig.ac.jp/whatsnew/2008/080606-e.html

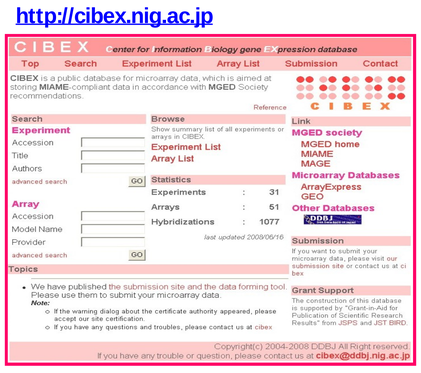
Gene Expression Data

CIBEX

Center for Information Biology gene EXpression database

- Public repository for microarray data
- Curated resource for data browsing and download

http://cibex.nig.ac.jp



Raw Read Data

Trace/Short Read Archives

Trace Archive (TA) and Short Read Archive (SRA)

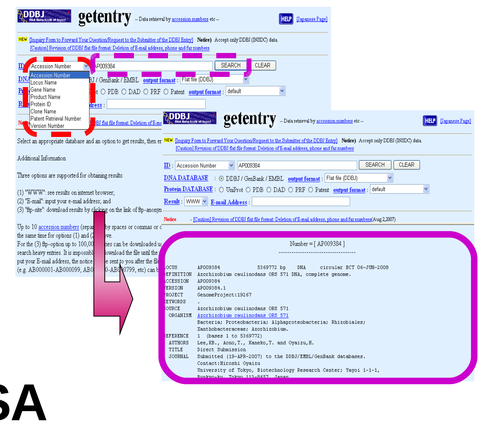
http://www.ddbj.nig.ac.jp/sub/trace_sra-e.html

"Next Inflation" of Data Submissions Would Be Short Reads Outputted from 454, Solexa, SOLiD, etc.

getentry

Data retrieval by accession numbers, etc.

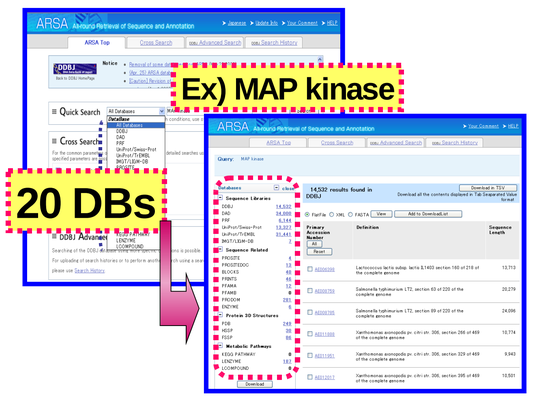
http://getentry.ddbj.nig.ac.jp/top-e.html



ARSA

All-round Retrieval of Sequence and Annotation

http://arsa.ddbj.nig.ac.jp/top-e.html



WABI (Web API for Biology)

Web Application Program Interface

http://www.xml.nig.ac.jp

Homology Search

FASTA | BLAST | PSI-BLAST | SSEARCH

http://www.ddbj.nig.ac.jp/search/top-e.html

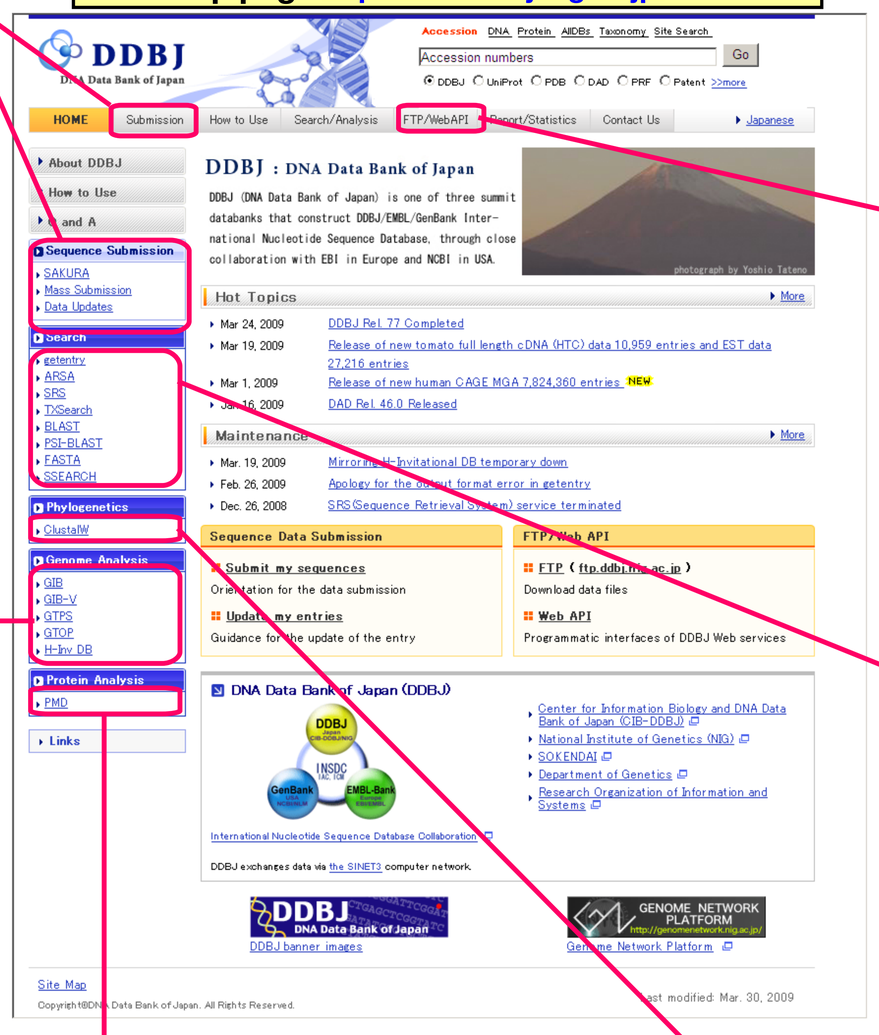
HMMPFAM

A motif search program against Pfam on the basis of Hidden Markov Model (HMM)

http://hmmpfam.ddbj.nig.ac.jp/top-e.html

Submission to DDBJ

DDBJ top page http://www.ddbj.nig.ac.jp/

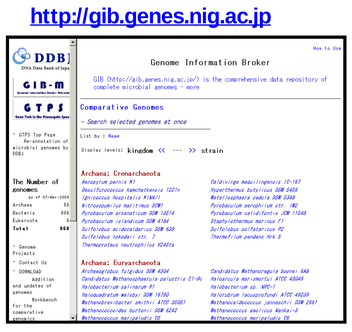


GIB

Genome Information Broker

http://gib.genes.nig.ac.jp

Comprehensive data repository of complete microbial genomes in the public domain

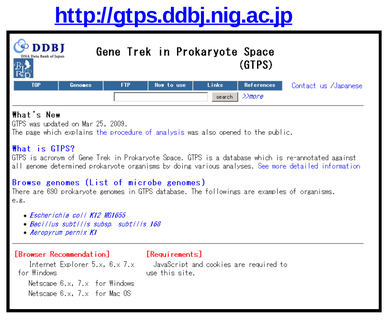


GTPS

Gene Trek in Prokaryote Space

http://gtps.ddbj.nig.ac.jp

Reannotated GIB data by categorized grade



H-InvDB (mirroring)

Database of full-length cDNAs assigned functional annotation as a result of H-Invitational

http://hinvdb.ddbj.nig.ac.jp/ahg-db/index.jsp

Genome Analyses

Database Search

Structure and Function

GTOP

Genome to protein structure and function

http://spock.genes.nig.ac.jp/~genome/gtop.html

PMD

Protein Mutant Database

http://pmd.ddbj.nig.ac.jp/~pmd/pmd.html

Phylogenetics

ClustalW

Multiple alignment

http://clustalw.ddbj.nig.ac.jp/top-e.html

TXSearch

Retrieval of unified taxonomy database

http://txsearch.ddbj.nig.ac.jp/top-e.html