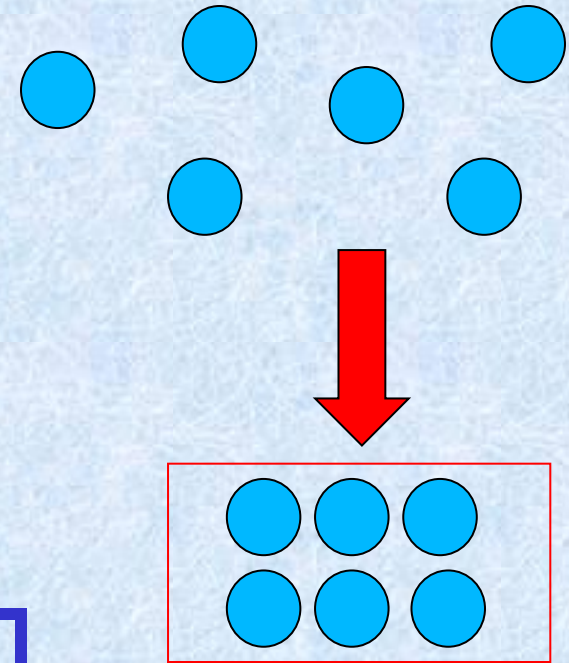# Databases

Biological experiments **(macromolecular sequences, structures, expression profiles, pathways etc)** provide wealth of data.

The data are available randomly in the literature

It is necessary to collect the scattered data and put in proper order in the form of a database

Database is an organized collection of information, in computer-readable form.

Databases: Characteristics

1. The contents

2. The ontology: the list of valid terms and their definitions

3. The logical structure, or the expression of the inter-relationships among the data, called the schema.

4. The format of the data

5. The routes for selective retrieval of data, and presentation of results, or pasting them on to a program for analysis

6. Links to other information resources: other databases, references to original publications of data, tutorial background etc.

Characteristics

Panel 1

NO.	193
***** Sequence and structural information*****	
PROTEIN NAME	Lysozyme
SOURCE	Bacteriophage T4
LENGTH	164
MOLECULAR WEIGHT	18603.94
PIR_ID	LZBPT4
SWISSPROT_ID	LYCV_BPT4 (P00720)
E.C.NUMBER	EC3.2.1.17 Go to BRENDA
PMD.NO	A921426
PDB_wild	2LZM Homologous PDB Entries
PDB_mutant	1L56
MUTATION	K 60 P
NO. OF MOLECULE	1
SECONDARY STRUCTURE	Helix
ACCESSIBLE SURFACE AREA	107.7 A**2

1

Panel 4

```
NO.
PROTEIN           Staphylococcal nuclease
SOURCE            Staphylococcus aureus
MUTATION           G 20 E
MUTATED_CHAIN
PDB_mutant
SEC.STR.
ASA
NO_MOLECULE
STATE
dG_H2O        6.3
ddG_H2O      -3.2
dG
ddG
T              20
Tm
dTm
dHvH
dHcal
m             4.7
Cm            1.3
dCp
pH            3.9
BUFFER_NAME     HEPES
BUFFER_CONC     10 mM
ION_NAME_1      NaCl
ION_CONC_1      100 mM
ADDITIVES
PROTEIN_CONC    50 microg/mL
MEASURE         F1
METHOD          GdnHCl
REVERSIBILITY   unknown
ACTIVITY
ACTIVITY_Km
ACTIVITY_Kcat
ACTIVITY_Kd
KEY_WORDS       dielectric effect; electrostatics; hydration; pKa; bioenergetics
REFERENCE       Proc Natl Acad Sci U S A. 2010 Sep 14;107(37):16096-100.
AUTHOR
REMARKS         Isom DG, Castañeda CA, Cannon BR, Velu PD, Garcú)-Moreno E B.
//
```

4

Panel 5

Entry	[ ] - [ ]	PDB Code	[ ]	Start Clear
Protein	Lysozyme	Source	Bacteriophage	
Mol-weight	[ ] To [ ]			
Mutation	[ ] To [ ] ☑ Single ☐ Double ☐ Multiple ☐ Wild Type			
Sec.Structure	☑ Helix ☑ Sheet ☐ Turn ☐ Coil			
Accessibility	☐ Any ☐ Buried ☐ Partially Buried ☐ Exposed ◉ ASA 0 To 20 %			
Measure	☐ Absorbance ☐ CD ☐ DSC ☐ Fluorescence ☐ NMR ☐ Others			
Method	☑ Thermal ☐ Denaturants ☐ Others			
pH	5 To 9			
dTm/Tm/T	dTm ▾ 0 To 50 C ▾			
dH/dCp/dG/dG_H2O	dH ▾ [ ] To [ ] energy unit: kcal ▾			
ddG/ddG_H2O	ddG ▾ [ ] To [ ]			
State	☑ 2 ☐ 3 ☐ >3			
Reversibility	Yes ▾			
Keyword	[ ] OR ▾ [ ]			
Author	[ ] OR ▾ [ ]			
Year	Since [ ] Until [ ]			

5

Panel 2

Terms	Explanations
No	Entry number. This option can be used for getting data from a particular entry (Eg. 3012) or a range of entries (Eg. 10107-10365) with/without other search conditions
Protein Name	Name of the protein. Multiple words can be entered with spaces. Wild card can also be used: * for a string of characters and ? for a character. If * and ? were to be used as real characters, place "\" (backslash) before them.
E.C.No.	Enzyme Commission number
PMD No.	Protein Mutant Database accession number
PDB_wild	Protein Data Bank code for the native protein

2

Panel 3

Thermodynamic data
$\Delta G, \Delta H, T_m, \Delta C_p, \Delta G^{H2O}, m, C_m,$ etc.

Methods, Condition
T, pH, buffer, ion
protein concentration

Literature information
Reference, authors

collection

3DinSight integrated database

Sequence PIR, SwissProt

Structure PDB

Function Motif, Mutation

Other databases EC, BRENDA, PubMed

ProTherm

Analysis

Search, Visualization

Interface

Internet

3

Panel 6
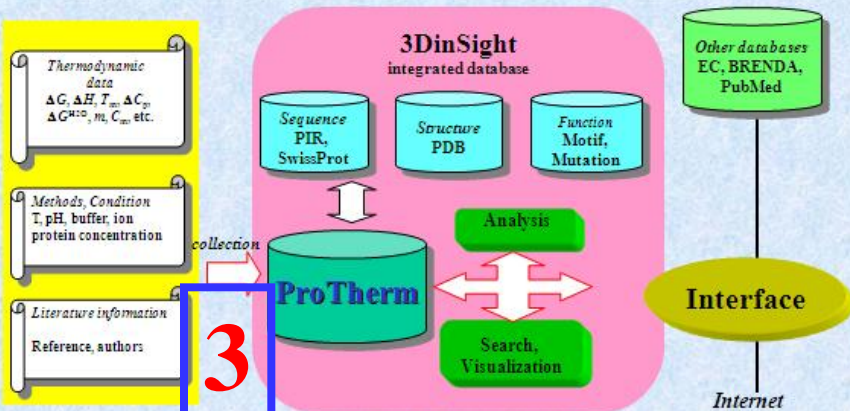
Entry	Protein	PDB_wild	PDB_mutant	Mutation	Sec.Str.	ASA (%)	Tm	dTm	pH	REFERENCE
1459	LYSOZYME	2LZM	1QT6	E 11 H	H	18.89	65.20	0.10	5.40	PROC NATL ACAD SCI U S A 92, 452-456 (1995)
1522	LYSOZYME	2LZM	1L87	F 153 L	H		0.05	65.68	0.80	5.70	J MOL BIOL 229, 747-769 (1993)
199	LYSOZYME	2LZM	1L23	G 77 A	H	4.83	65.60	0.90	6.50	BIOPOLYMERS 32, 1431-1441 (1992)

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health

Search: PubMed   Limits Advanced search Help

Display Settings: ☑ Abstract

Nucleic Acids Res. 1999 Jan 1;27(1):286-8.

ProTherm: Thermodynamic Database for Proteins and Mutants.

Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A.

Tsukuba Life Science Center, The Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

6

Organization: Relational database

In 1970, E.F. Codd from IBM described the relational database.

The basic unit of a relational database is a set of correspondence between different features of the database contents, called tables.

Relational database is the one in which data are organized as tables, each table comprising a group of records with the same fields (known as attributes). This allows related data to be linked (reassembled) as required without reorganizing the original tables.

The set theoretic operations (union, intersection, difference, Cartesian product) on tables facilitate processing of logically complex queries.

Example

Table 1

Two tables from a relational database of properties of amino acids

Amino acid	3-letter code	1-letter code	Volume (Å³)	Surface area (Å²)	Distal group
Alanine	Ala	A	88.6	115	Methyl
Arginine	Arg	R	173.4	225	Guanidinium
Asparagine	Asn	N	111.1	150	Amide
Aspartic acid	Asp	D	114.1	160	Carboxyl
Cysteine	Cys	C	108.5	135	Sulphydryl
Glutamic acid	Glu	E	138.4	190	Carboxyl
Glutamine	Gln	Q	143.8	180	Amide
Glycine	Gly	G	60.1	75	Hydrogen
Histidine	His	H	153.2	195	Imidazole
Isoleucine	Ile	I	166.7	175	Methyl
Leucine	Leu	L	166.7	170	Methyl
Lysine	Lys	K	168.6	200	Amino
Methionine	Met	M	162.9	185	Methyl
Phenylalanine	Phe	F	189.9	210	Phenyl
Proline	Pro	P	112.7	145	Pyrrolidine
Serine	Ser	S	89.0	115	Hydroxyl
Threonine	Thr	T	116.1	140	Hydroxyl
Tryptophan	Trp	W	227.8	255	Indole
Tyrosine	Tyr	Y	193.6	230	Phenol
Valine	Val	V	140.0	155	Methyl

Two tables from a relational database of properties of amino acids (continued)

Distal group	H-bond donor	H-bond acceptor
Amide	yes	yes
Amino	yes	no
Carboxyl	no	yes
Guanidinium	yes	yes
Hydrogen	no	no
Hydroxyl	yes	yes
Indole	yes	yes
Methyl	no	no
Phenol	yes	yes
Phenyl	no	no
Pyrrolidine	yes	no
Sulphydryl	yes	no

Simple: What are the three letter codes of the amino acids, which have distal carboxyl group?

View

Compound: What are the three letter codes of the amino acids with volume more than 125 A³ and have distal carboxyl group?

What are the three letter codes of the amino acids, which can serve as hydrogen bond donors?

Join

Lesk, 2008

M. Michael Gromiha, IITM, BT3040

Example

General form of joining is the Cartesian product of the two tables. If the set contains n and m elements the product will contain **nm** elements. Here, 20 amino acids and 12 distal groups and the total will be 240 rows.

From Table 1						**From Table 2**		
AA	**3**	**1**	**V**	**A**	**Group**	**Group**	**Donar**	**Acceptor**
Alanine	Ala	A	88.6	115	Methyl	Amide	Yes	Yes
Alanine	Ala	A	88.6	115	Methyl	Amide	Yes	No
Alanine	Ala	A	88.6	115	**Methyl**	**Methyl**	No	No
Aspartic acid	Asp	D	114.1	160	**Carboxyl**	**Carboxyl**	No	Yes

Three letter codes of amino acids that have side chains that could serve as hydrogen bond acceptors:

Natural join

Complex queries

What are the **three letter codes** of amino acids with **volumes** between 100 and 150 AND [(that can serve as **hydrogen bond donors** AND NOT serve as **hydrogen bond acceptors**) OR (that have **surface areas** greater than 120 A2 AND have **distal methyl groups**)].

The **structured Query Language (SQL) is** fairly well standardized syntax for probing relational databases with complex queries.

Complex queries containing logical connectivities are translatable into Codd's set of operations on tables.

Syntax

SELECT <3_letter_code> from <amino_acid _table)

WHERE (sidechain_volume between 100 and 150)

AND

(H-bond_donor = "yes" AND H-hond_acceptor = "no")

OR

(surface_area > 120 AND distal_group = "methyl"))

Database collections

Nucleic acid research Database issue (First issue in every year). It is available for free access.

http://nar.oupjournals.org/

Listing of databases

http://www.oxfordjournals.org/nar/database/a/

Database categories

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Thermodynamic databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases (PUBMED)
- Organelle databases
- Plant databases
- Immunological databases

Nucleotide sequence databases

International collaboration

1. DDBJ (DNA Data Bank of Japan)
2. EMBL (European Molecular Biology Laboratory)
3. Genbank (USA)
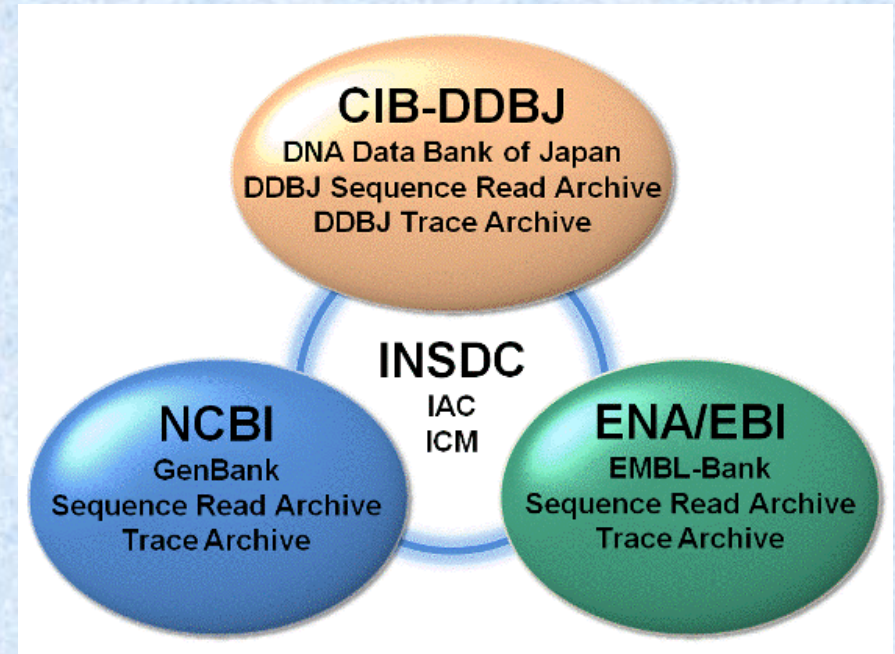
They exchange sequences via SINET3 Computer network
SINET: Science Information Network

1. DNA Data Bank of Japan

http://www.ddbj.nig.ac.jp/
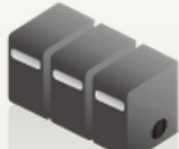
DDBJ: Growth
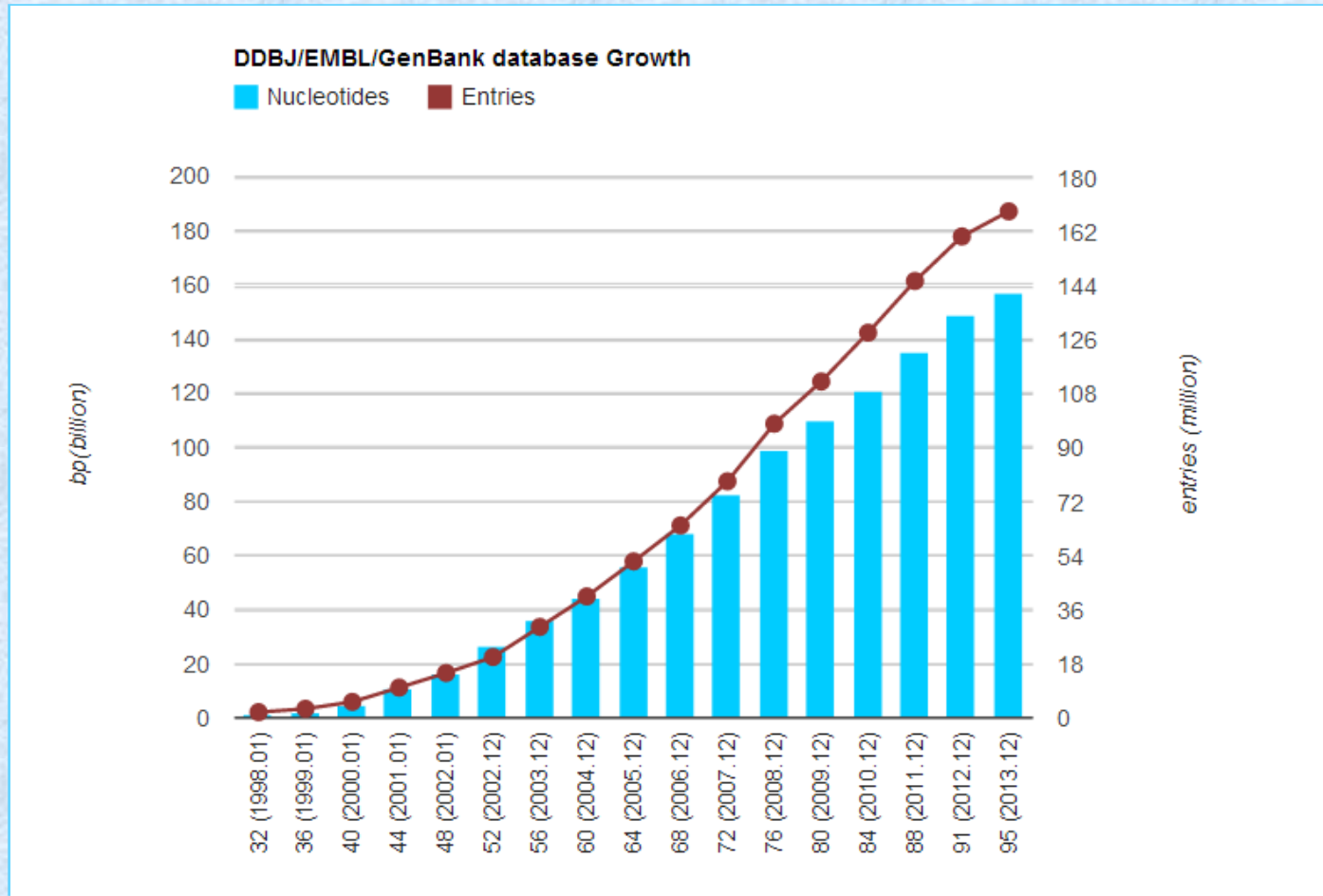
Dec 2016: > 2.1 Trillion nucleotides
Number of sequences: > 790 millions

DDBJ: Topmost 10 organisms

No. Organism	Nucleotides	Entries
001 Homo sapiens	17382318654 bp	20522483
002 Mus musculus	9988986985 bp	9727377
003 Rattus norvegicus	6525616319 bp	2197920
004 Bos taurus	5390887315 bp	2202851
005 Zea mays	5076664404 bp	3963159
006 Sus scrofa	4890614452 bp	3289376
007 Danio rerio	3120659920 bp	1726789
008 Marine metagenome	2482805950 bp	3173890
009 Vitis vinifera	1555395119 bp	810152
010 Hordeum vulgare	1455338435 bp	1009480

Search

Contents

Name
Source
Accession number
Keywords
Authors
Reference
PUBMED index
Nucleotide sequence
Number of bases
 A, T, C and G
Protein sequence (translated)

```
LOCUS       Z36833                  315 bp    mRNA     linear   HUM 10-APR-1997
DEFINITION  H.sapiens (xs4) mRNA, 315bp.
ACCESSION   Z36833
VERSION     Z36833.1
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 315)
  AUTHORS   Mueller-Pillasch,F., Gress,T., Lehrach,H. and  Adler,G.
  TITLE     Differential gene expression in pancreatic cancer. Use of an
            automated approach for the large scale isolation and
            characterisation of cDNA clones containing differentially expressed
            sequences.
  JOURNAL   Unpublished.
REFERENCE   2  (bases 1 to 315)
  AUTHORS   Gress,T.
  JOURNAL   Submitted (16-AUG-1994) to the INSDC. Gress T., University of Ulm,
            Department of Internal Medicine I, Robert Koch Str.8, 89081 Ulm,
            Germany, 89081
REFERENCE   4  (bases 1 to 315)
  AUTHORS   Gress,T.M., Muller-Pillasch,F., Geng,M., Zimmerhackl,F.,
            Zehetner,G., Friess,H., Buchler,M., Adler,G. and  Lehrach,H.
  TITLE     A pancreatic cancer-specific expression profile
  JOURNAL   Oncogene 13(8), 1819-1830(1996).
  PUBMED    8895530
FE
     BASE COUNT         80 a         74 c         82 g         76 t
     ORIGIN
        1 atcagcggat ttgcgtcgta ttgggcgcct ggatcaccag ggctgctttt atctctggta
       61 aagtggatat tgttgacatc actgacccc acattgacca catatacatg gtttacatgt
      121 tccaatatga ctccacccat gagatattcn atgacaccga caggggtgag aacgggnagc
      181 ttgacatcaa tggaaatccc acaccatctn cgaggagaga catcctccaa catcatgtgg
      241 cgagatgttg cgctgagtac gtcgtggagt cactgtgtct cacacatgag aggtgtgctc
      301 atggagggggg agcaa
     //
```

Advanced search

Quick Search

Field   Show examples

Primary Accession Number [_____]

Search Result

▶ Facet

▶ List of Entries

1 - 30 entries / Number of founds: 26506   ⦿ FlatFile

PrimaryAccessionNumber	Definition	Se
☐ AF217656	Definition:Human papillomavirus is	
	Organism:Human papillomavirus	
☐ AF217657	Definition:Human papillomavirus is	
	Organism:Human papillomavirus	
☐ AF217659	Definition:Human papillomavirus is	
	Organism:Human papillomavirus	
☐ AF217658	Definition:Human papillomavirus	
	Organism:Human papillomavirus	
☐ AF217660	Definition:Human papillomavirus	
	Organism:Human papillomavirus	
☐ AF217661	Definition:Human papillomavirus	
	Organism:Human papillomavirus	
☐ AF455142	Definition:Human papillomavirus	
	Organism:Human papillomavirus	
☐ AF455144	Definition:Human papillomavirus	
	Organism:Human papillomavirus	
☐ AF455146	Definition:Human papillomavirus	
	Organism:Human papillomavirus	

Reference PubmedID [_____]

```
LOCUS       AF217656                434 bp    DNA     linear   VRL 30-NOV-2000
DEFINITION  Human papillomavirus isolate FA14 major capsid protein L1 gene,
            partial cds.
ACCESSION   AF217656
VERSION     AF217656.1
KEYWORDS    .
SOURCE      Human papillomavirus
  ORGANISM  Human papillomavirus
            Viruses; dsDNA viruses, no RNA stage; Papillomaviridae;
            unclassified Papillomaviridae.
REFERENCE   1  (bases 1 to 434)
  AUTHORS   Antonsson,A., Forslund,O., Ekberg,H., Sterner,G. and Hansson,B.G.
  TITLE     The ubiquity and impressive genomic diversity of human skin
```

```
            /translation="NIYNNQGTRLEVPKVSGNQHRVFRLKLPDPNRFALADMSVYNPD
            KERLVWGLKGIEIGRGQPLGIGSSGHPLFNKVNDTENGNTYRNSSKDDRQNISFDPKQ
            LQMFIIGCTPCIGEHWDRAPACVNDDQAGRCPPIELINSYIQ"
BASE COUNT        153 a           71 c           93 g          117 t
ORIGIN
        1 tcaatattta taacaatcaa ggcacacgat tggaggttcc taaagtatca ggaaatcaac
       61 acagggtatt tagattaaag ctaccagatc ctaataggtt tgcgttagct gacatgtcag
      121 tatataaccc tgacaaagaa agattagtat ggggtttgaa aggcatagaa ataggcaggg
      181 gccaaccttt aggaataggc agcagtggtc atccactgtt taataaggtt aatgatacag
      241 aaaatggcaa tacatatagg aactcctcta aggatgatag acaaaatatt tcatttgacc
      301 ccaagcagtt gcaaatgttt attattggct gtactccatg tataggagaa cattgggaca
      361 gagcaccagc atgtgttaat gatgatcaag ctggtagatg tcctcctata gagttaataa
      421 actcatatat acag
//
```

DDBJ Data Submission

- Single sequence
- Multiple sequences
- Updates

Sequence, address, contact details, status of publications.

SAKURA

SAKURA is a nucleotide sequence data submission system through the WWW server at DDBJ.
SAKURA has been open to public and continuously refined since 1995. Using this system, you can interactively enter and submit nucleotide and translated amino acid sequences, functions and features of the sequences, and references as well as your name, affiliation and address.

Mass Submission System (MSS)
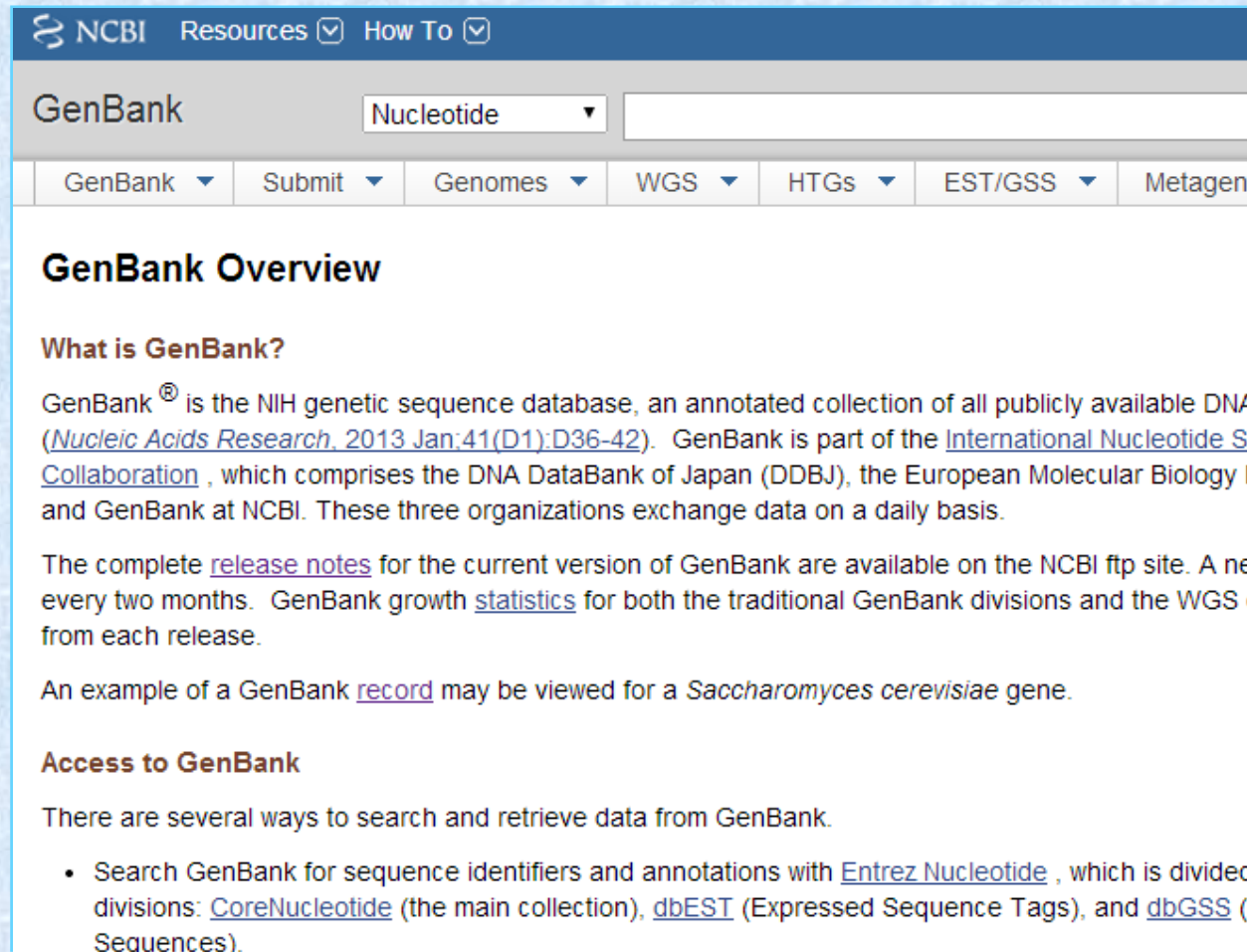
Japanese

Go

We recommend using Mass Submission System (MSS) when:

1. The submission consists of large number of entries.
2. The submission involves long nucleotide, complex submission resulting in a many features such as genome data.
3. The submission is unsuitable for SAKURA

Genbank

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences

Genbank: contents

```
LOCUS       SCU49845      5028 bp    DNA            PLN        21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrenc
  TITLE     Cloning and sequence of REV7, a gene whose fun
            DNA damage-induced mutagenesis in Saccharomyce
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requi
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology,
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
```

```
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaatttca cttttcatat gagaacttat tttcttattc
```

Genbank: contents

The LOCUS field contains locus name, sequence length, molecule type, GenBank division, and modification date

Definition: Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function

Accession: The unique identifier for a sequence record

Version: A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. GI: GenInfo identifier.

Keywords: Word or phrase describing the sequence

Source: organism name

Reference: Publications by the authors of the sequence that discuss the data reported in the record.

The GenBank database is divided into 18 divisions:
1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences

Genbank: contents

Title: TMFunction: database for functional residues in membrane proteins.

Authors: Gromiha MM, Yabuki Y, Suresh MX, Thangakani AM, Suwa M, Fukui K.

Journal: Nucleic Acids Res. 2009 Jan;37(Database issue):D201-4.

PMID: 18842639

Features: Information about genes and gene products, as well as regions of biological significance reported in the sequence.

CDs: Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons).

complete feature is simply written as *n..m*
Example: 687..3158

<   indicates **partial on the 5' end**
Example: <1..206

>   indicates **partial on the 3' end**
Example: 4821..5028>

(complement) indicates that the feature is on the complementary strand
Example: complement(3300..4037)
The feature extends from base 3300 through base 4037 but is actually on the complementary strand.

Translation: The amino acid translation corresponding to the nucleotide coding sequence

Base count: Frequency of occurrence of the different base types, A, C, G and T in the sequence.

http://www.ncbi.nlm.nih.gov/genbank/

Genbank: search

EMBL: Nucleotide sequence database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are <u>direct submissions</u> from individual researchers, genome sequencing projects and patent applications.

http://www.ebi.ac.uk/embl/

EMBL: Statistics

Release on 16-Jan-2014.

The release contains 798.2 million sequence entries comprising 2004.9 billion nucleotides.
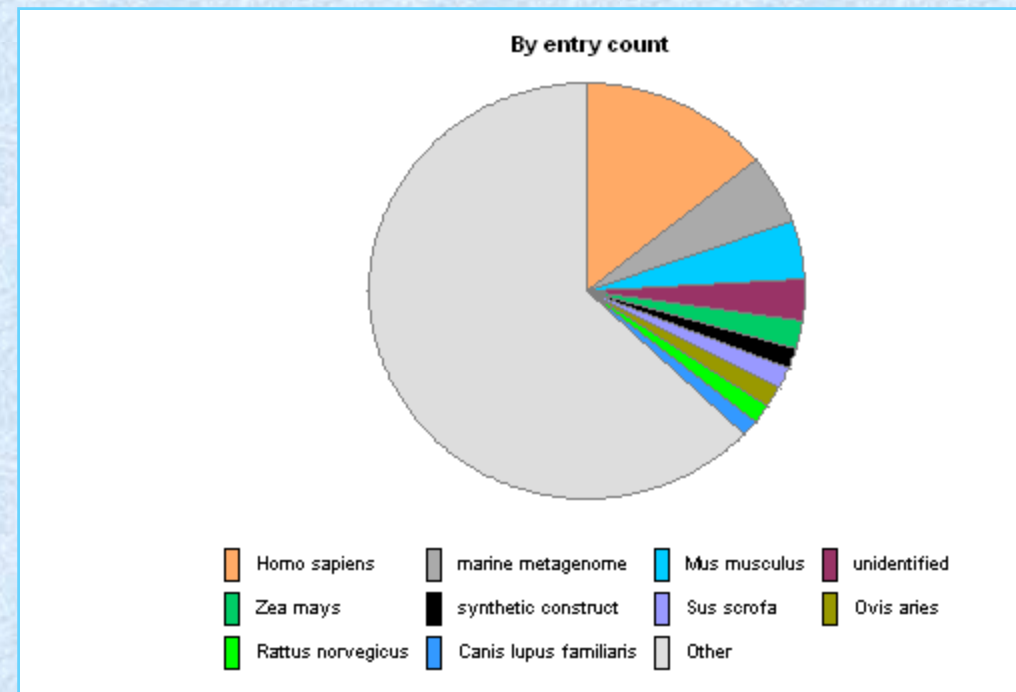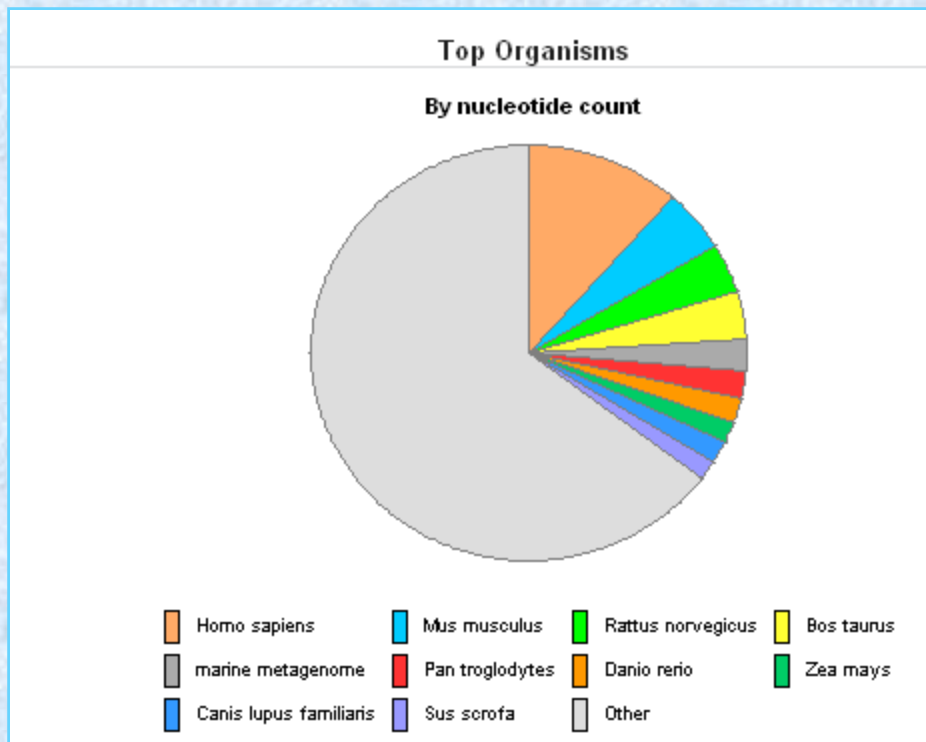
Database entry

Databases | Tools | Research | Training | Indust

EBI > Databases > Database Browsing > Dbfetch > EMBL-Bank: TRBG361

EBI Dbfetch

```
ID    X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC    X56734; S46826;
XX
DT    12-SEP-1991 (Rel. 29, Created)
DT    25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE    Trifolium repens mRNA for non-cyanogenic beta-glu
XX
KW    beta-glucosidase.
XX
OS    Trifolium repens (white clover)
OC    Eukaryota; Viridiplantae; Streptophyta; Embryoph
OC    Spermatophyta; Magnoliophyta; eudicotyledons; co
OC    fabids; Fabales; Fabaceae; Papilionoideae; Trifo
XX
RN    [5]
RP    1-1859
RX    DOI; 10.1007/BF00039495
RX    PUBMED; 1907511.
RA    Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT    "Nucleotide and derived amino acid sequence of th
RT    beta-glucosidase (linamarase) from white clover
RL    Plant Mol. Biol. 17(2):209-219(1991).
XX
RN    [6]
RP    1-1859
RA    Hughes M.A.;
RT    ;
RL    Submitted (19-NOV-1990) to the INSDC.
```

```
XX
FH    Key             Location/Qualifiers
FH
FT    source          1..1859
FT                    /organism="Trifolium repens"
FT                    /mol_type="mRNA"
FT                    /clone_lib="lambda gt10"
FT                    /clone="TRE361"
FT                    /tissue_type="leaves"
FT                    /db_xref="taxon:3899"
FT    CDS             14..1495
FT                    /product="beta-glucosidase"
FT                    /EC_number="3.2.1.21"
FT                    /note="non-cyanogenic"
FT                    /db_xref="GOA:P26204"
FT                    /db_xref="InterPro:IPR001360"
FT                    /db_xref="InterPro:IPR013781"
FT                    /db_xref="InterPro:IPR017853"
FT                    /db_xref="InterPro:IPR018120"
FT                    /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                    /protein_id="CAA40058.1"
FT                    /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                    FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                    DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                    VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                    CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                    DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                    IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                    EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                    IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT    mRNA            1..1859
FT                    /experiment="experimental evidence, no additional details
FT                    recorded"
XX
SQ    Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
      aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
      cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
      tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
      aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
```

EMBL: Search

```
ID     M33197; SV 1; linear; mRNA; STD; HU      FT    mRNA              <1..1268
XX                                              FT                      /note="GAPDH mRNA"
AC     M33197;                                  FT    CDS               61..1068
XX                                              FT                      /codon_start=1
DT     10-JUL-1990 (Rel. 24, Created)           FT                      /gene="GAPD"
DT     17-APR-2005 (Rel. 83, Last updated,      FT                      /note="glyceraldehyde-3-phosphate dehydrogenase (EC
XX                                              FT                      1.2.1.12)"
DE     Human glyceraldehyde-3-phosphate de      FT                      /db_xref="GDB:119249"
XX                                              FT                      /db_xref="GOA:P04406"
KW     glyceraldehyde-3-phosphate dehydrog      FT                      /db_xref="H-InvDB:HIT000195501"
XX                                              FT                      /db_xref="HGNC:4141"
OS     Homo sapiens (human)                     FT                      /db_xref="InterPro:IPR006424"
OC     Eukaryota; Metazoa; Chordata; Cran       FT                      /db_xref="InterPro:IPR016040"
OC     Eutheria; Euarchontoglires; Primate      FT                      /db_xref="InterPro:IPR020828"
OC     Homo.                                    FT                      /db_xref="InterPro:IPR020829"
XX                                              FT                      /db_xref="InterPro:IPR020830"
RN     [1]                                      FT                      /db_xref="InterPro:IPR020831"
RP     1-1268                                   FT                      /db_xref="PDB:1U8F"
RX     PUBMED; 3664468.                         FT                      /db_xref="PDB:1ZNQ"
RA     Tokunaga K., Nakamura Y., Sakata K       FT                      /db_xref="PDB:2FEH"
RA     Sakiyama S.;                             FT                      /db_xref="PDB:3GPD"
RT     "Enhanced expression of a glyceral       FT                      /db_xref="UniProtKB/Swiss-Prot:P04406"
RT     human lung cancers";                     FT                      /protein_id="AAA52518.1"
RL     Cancer Res. 47(21):5616-5619(1987)       FT                      /translation="MGKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVY
XX                                              FT                      MFQYDSTHGKFHGTVKAENGKLVINGNPITIFQERDPSKIKWGDAGAEYVVESTGVFTT
DR     Ensembl-Gn; ENSG00000111640; Homo_s      FT                      MEKAGAHLQGGAKRVIISAPSADAPMFVMGVNHEKYDNSLKIISNASCTTNCLAPLAKV
DR     Ensembl-Tr; ENST00000229239; Homo_s      FT                      IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNIIPASTGAAKAVGKVIP
DR     Ensembl-Tr; ENST00000396859; Homo_s      FT                      ELNGKLTGMAFRVPTANVSVVDLTCRLEKPAKYDDIKKVVKQASEGPLKGILGYTEHQV
DR     Ensembl-Tr; ENST00000396861; Homo_s      FT                      VSSDFNSDTHSSTFDAGAGIALNDHFVKLISWYDNEFGYSNRVVDLMAHMASKE"
XX                                              XX
FH     Key              Location/Qualifiers     SQ    Sequence 1268 BP; 295 A; 385 C; 326 G; 262 T; 0 other;
FH                                                    gttcgacagt cagccgcatc ttcttttgcg tcgccagccg agccacatcg ctcagacacc      60
FT     source           1..1268                       atgggaagg tgaaggtcgg agtcaacgga tttggtcgta ttgggcgcct ggtcaccagg      120
FT                      /organism="Homo sap            gctgctttta actctggtaa agtggatatt gttgccatca atgacccctt cattgacctc      180
FT                      /map="12p13"                   aactacatgg tttacatgtt ccaatatgat tccacccatg gcaaattcca tggcaccgtc      240
FT                      /mol_type="mRNA"               aaggctgaga acgggaagct tgtcatcaat ggaaatccca tcaccatctt ccaggagcga      300
```

Literature database

Chemical abstracts

Physical abstracts

Science citation index

PUBMED

Mainly for biological sciences and are freely available

Useful for getting the references of any work and related papers

PUBMED

Search with the keyword "DDBJ Nucleic acid sequence database"

http://www.ncbi.nlm.nih.gov/pubmed/

Abstract

What is Open Access?

Full text

Nucleic Acids Research

Formats: Abstract | Full Text | PDF (3.6M)

DDBJ progress report

Eli Kaminuma, Takehide Kosuge, Yuichi Kodama, Hideo Aono, Jun Mashima, Takashi Gojobori, Hideaki Sugawara, Osamu Ogasawara, Toshihisa Takagi, Kousaku Okubo, and Yasukazu Nakamura[*]

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima, 411-8510, Japan
*To whom correspondence should be addressed. Tel: Phone: +81 55 981 6859; Fax: +81 55 981 6889; Email: yanakamu@genes.nig.ac.jp

ABSTRACT   Other Sections▼

The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval and annotation analysis. The DDBJ collected and released 3 637 446 entries/2 272 231 889 bases between July 2009 and June 2010. A highlight of the released data was archive datasets from next-generation sequencing

PubMed articles by these authors

- Kaminuma, E.
- Kosuge, T.
- Kodama, Y.
- Nakamura, Y.

PubMed related articles

- DDBJ launches a new archive database with analytical tools for next-generation sequence data. [Nucleic Acids Res. 2010]
- Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technolo [Adv Exp Med Biol. 2010]
- The sequence read archive. [Nucleic Acids Res. 2011]
- *Review* The EMBL Nucleotide Sequence Database. Contributing and accessing data. [Mol Biotechnol. 1999]
- *Review* From mapping to sequencing, post-sequencing and beyond. [Plant Cell Physiol. 2005]

  » See reviews... | » See all...

Recent Activity   Turn Off   Clear

DDBJ progress report

Related articles

PUBMED: Features

Search with keywords display the results from the latest article.

Search with "related entries" show the output from the most relevant article.

One can also use the mesh terms to get a specific article. E.g. Nucleic Acids Res. 34, D204-D206 (2006)

204-206[PAGE] AND 34[VOL] AND 2006[PDAT]

Disease database

Alterations in proteins cause cancer.

Dr. Akinoi Sarai's group in Japan

Protein function database

Search Results

Important residues for protein function
M.M. Gromiha et al.
CBRC, AIST, Japan

M.M. Gromiha et al. (2009) Nucleic Acids Res. 37, D201-204

TMFunction: database for functional residues in membrane proteins

M. Michael Gromiha[1,*], Yukimitsu Yabuki[1], M. Xavier Suresh[1], A. Mary Thangakani[2], Makiko Suwa[1] and Kazuhiko Fukui[1]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064 and [2]Advanced Technology Inc., Tokyo, Japan