

PR15_Genomske in proteomske podatkovne baze

Biomolekularne podatkovne baze

Examples of biomolecular databases

- **Sequence and structure databases**
 - Protein sequences (UniProt)
 - DNA sequences (EMBL, Genbank, DDBJ)
 - 3D structures (PDB)
 - Structural motifs (CATH)
 - Sequence motifs (PROSITE, PRODOM)
- **Genome sequences and annotations**
 - Genome-specific databases (SGD, FlyBase, AceDB, PlasmODB, ...)
 - Multiple genomes (Integr8, NCBI, KEGG, TIGR, ...)
- **Molecular functions**
 - Transcriptional regulation (TRANSFAC, RegulonDB, InteractDB)
 - Enzymatic catalysis (Expasy, LIGAND/KEGG, BRENDA)
 - Transport (YTPdb)
- **Biological processes**
 - Metabolic pathways (EcoCyc, LIGAND/KEGG, Biocatalysis/biodegradation)
 - Signal transduction pathways (CSNdb, Transpath)
 - Protein-protein interactions (DIP, BIND, MINT)
 - Gene networks (GeneNet, FlyNets)

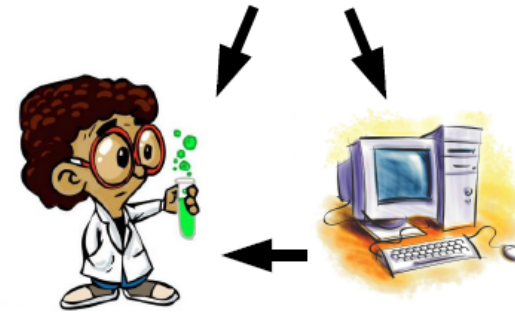
Why databases ?

- biology has turned into **data-rich science**
 - High-throughput genomics, proteomics, metabolomics, ...
 - Vast amount of data generated in experiments (like MS peptide fragments)
- need for storing and communicating large datasets has grown tremendously
 - archiving, curation, analysis and interpretation of all of these datasets are a challenge
 - convenient methods for proper storing, searching & retrieving necessary
- **Databases are the means** to handle this data overload



What can databases do ?

- **Make biological data available ...**
 1. ... to scientists.
 2. ... in computer-readable form.
 - Analysis (computer based)
 - Handle and share large volumes of data
 - Interface for computer based systems (Algorithms, Web interfaces)
- **Store data**
 - Defined formats
 - Automated storage and retrieval of experimental data
- **Link knowledge with external resources**



Database classification I

- **Type of data**
 - Nucleotide or protein sequences
 - Protein sequence patterns and motifs
 - Macromolecular 3D structures
 - Gene expression data
 - Metabolic pathways
 - ...
- **Data entry and quality control**
 - Scientists deposit data directly
 - Appointed curators add and update
 - Type and degree of error checking
 - Consistency, redundancy, conflicts, updates



Database classification II

- **Primary or derived data**
 - Primary: experimental results directly into database
 - Secondary: results of analysis of primary databases
- **Technical design**
 - Flat-files
 - Relational database (SQL)
 - Object-oriented database
 - Exchange/publication technologies
(FTP, HTML, COBRA, XML, SOAP)
- **Maintainer status**
 - Large, public institution funded by government (EMBL, NCBI)
 - Academic group or scientist
 - Commercial company



Sequence databases - Primary Databases



National Center for Biotechnology

GenBank® : NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. <http://www.ncbi.nlm.nih.gov/genbank/>



European Nucleotide Archive

ENA - The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing. <http://www.ebi.ac.uk/ena/>



DNA Data Bank of Japan

DDBJ - DNA Data Bank of Japan was established 1986. <http://www.ddbj.nig.ac.jp/>



INSDC - The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between DDBJ, ENA, and GenBank for over 18 years. <http://insdc.org/>

Searching sequence databases - Secondary Databases



Ensembl? - Ensembl is a joint project between EMBL - EBI and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. www.ensembl.org/index.html



Genome Bioinformatics

UCSC Genome Browser - This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neanderthal projects.



The Biology and Genome of *C. elegans*.



The Zebrafish Model Organism Database



Comprehensive Microbial Resource

Comprehensive Microbial Resource




How to find my database ?

- Nucleic Acid Research offers database issue every year
- Database Journals
 - Database: The Journal of Biological Databases and Curation
- Database portals
 - DBD (database of biological database)
 - Pathguide
- Websearch
 - <http://imgtfy.com/>

PRINT ISSN: 0305-1048
ONLINE ISSN: 1362-4952


Nucleic Acids Research

VOLUME 38 DATABASE ISSUE JANUARY 1, 2010
www.nar.oxfordjournals.org



Downloaded from <http://nar.oxfordjournals.org/> at Springer Australia user on April 26, 2015

Now Open Access
No barriers to access – all articles freely available online

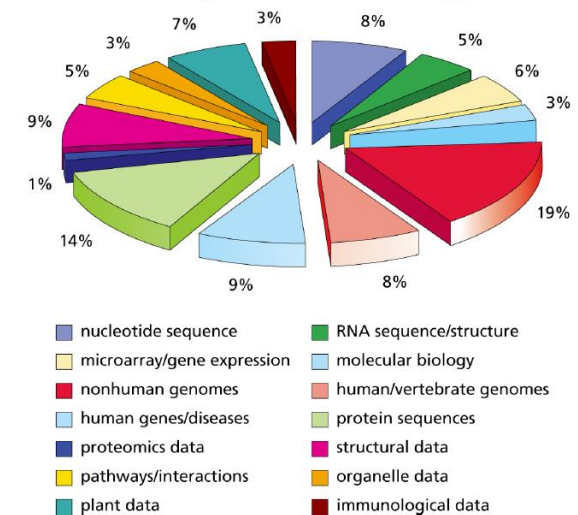
 OXFORD JOURNALS
A member of the OXFORD UNIVERSITY PRESS

Databases of databases

- There are hundreds of databases related to molecular biology and biochemistry. New databases are created every year.
- Every year, the first issue of Nucleic Acids Research is dedicated to biological databases
 - <http://nar.oupjournals.org/>
 - 2011 Issue: http://nar.oxfordjournals.org/content/39/suppl_1
- The same journal maintains a database of databases: the Molecular Biology Database Collection
 - <http://www.oxfordjournals.org/nar/database/c/>
- Some bioinformatics centres maintain multiple database, with cross-links between them. The SRS server at EBI holds an impressive collection of databases.
 - <http://srs.ebi.ac.uk/>

NAR Database Issue

- Online collection of biological databases:
<http://www.oxfordjournals.org/nar/database/c/>



Nucleic Acids Research

Oxford Journals > Life Sciences > Nucleic Acids Research > Database Summary Paper

2014 NAR Database Summary Paper

- ▶ [Compilation Paper](#)
- ▶ [Category List](#)
- ▶ [Alphabetical List](#)
- ▶ [Category/Paper List](#)
- ▶ [Search Summary Papers](#)

- [Nucleotide Sequence Databases](#)
- [RNA sequence databases](#)
- [Protein sequence databases](#)
- [Structure Databases](#)
- [Genomics Databases \(non-vertebrate\)](#)

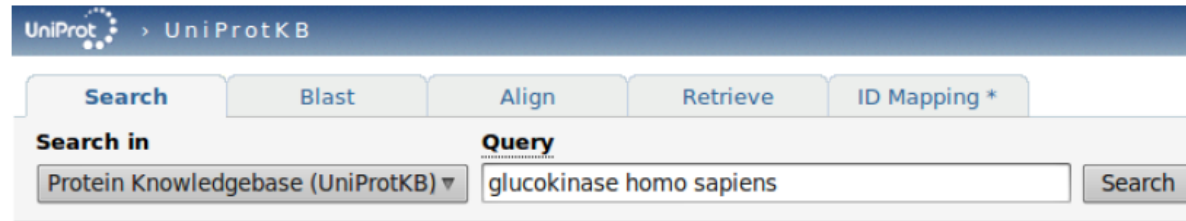
- [MGD - Mouse Genome Database](#)
- [The Gene Indices](#)
- [Genome annotation terms, ontologies and nomenclature](#)
- [Taxonomy and identification](#)
- [General genomics databases](#)

- [Animal Genome Size Database](#)
- [BaCMap](#)
- [Biodefense Proteomics Resource Center](#)
- [CAMERA](#)
- [COG - Clusters of Orthologous Groups of proteins](#)
- [CoGenT++](#)
- [diArk](#)
- [EBI Genomes](#)
- [Entrez Gene](#)
- [Entrez Genomes](#)
- [EPGD](#)
- [ERGO-Light](#)
- [GenDiS](#)
- [GeneNest](#)
- [GenoList](#)
- [Genome Project Database](#)
- [Genome Reviews](#)
- [GIB-IS - Genome information broker](#)
- [GOLD](#)
- [GtRDB - Genomic tRNA Database](#)
- [HOINVGEN](#)
- [Inparanoid](#)
- [Integr8 \(formerly Proteome Analysis Database\)](#)
- [KaryotypeDB](#)
- [KEGG - Kyoto Encyclopedia of Genes and Genomes](#)
- [MBGD - Microbial Genome Database](#)
- [MeGX](#)
- [MetaCyc](#)
- [Narcisse](#)
- [NegProt - Negative Proteome database](#)
- [NMPDR - National Microbial Pathogen Data Resource](#)
- [OrthoMCL](#)
- [ParameciumDB](#)
- [PartiGeneDB](#)
- [PEDANT](#)
- [PEP: Predictions for Entire Proteomes](#)
- [PhylomeDB](#)
- [STRING](#)
- [TDRtargets](#)
- [The Comprehensive Microbial Resource](#)
- [The Gene Indices](#)
- [TMBETA-GENOME](#)
- [TransportDB](#)

- [Viral genome databases](#)
- [Prokaryotic genome databases](#)
- [Unicellular eukaryotes genome databases](#)
- [Fungal genome databases](#)
- [Invertebrate genome databases](#)
- [Metabolic and Signaling Pathways](#)
- [Human and other Vertebrate Genomes](#)
- [Human Genes and Diseases](#)

How to access the data ?

- Human **Web interface** (web based, small scale)
 - Common mode of search are keywords with modifiers or identifiers
 - Cross-references link the information of different databases



The image shows a screenshot of the UniProtKB web interface. At the top, there is a blue header with the UniProt logo and the text 'UniProt > UniProtKB'. Below the header, there are five tabs: 'Search', 'Blast', 'Align', 'Retrieve', and 'ID Mapping *'. The 'Search' tab is selected. Underneath the tabs, there is a search form with two main sections: 'Search in' and 'Query'. The 'Search in' section has a dropdown menu currently set to 'Protein Knowledgebase (UniProtKB)'. The 'Query' section has a text input field containing the text 'glucokinase homo sapiens' and a 'Search' button to the right.

- **Web service** (SOAP, CORBA)
- **Flat files** (script based, large scale)
- **Database dump** (script based, large scale)

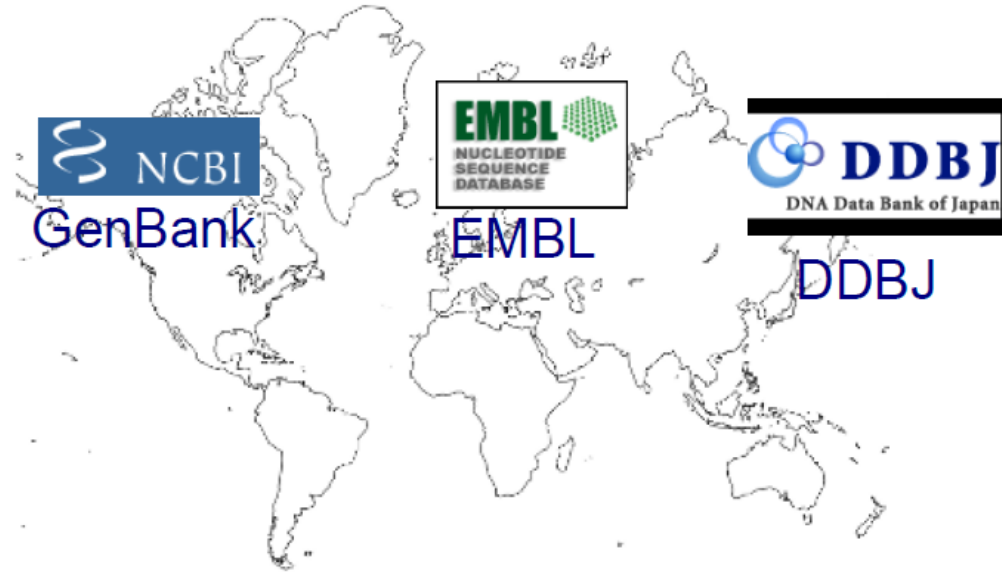
Database Tools

- Database design (Model building)
 - Determine the relationships between the different data elements.
 - Superimpose a logical structure upon the data on the basis of these relationships.
 - Scheme development (paper & pencil)
 - Scheme implementation and refinement (database designer like MicroOLAP DB Designer)
- Relational database (Storage)
 - MySQL, PostgreSQL, SQLite
- Interfaces (Access)
 - SQL queries
 - Administration tools (phpMySQL, phpPgAdmin)
 - Frameworks & Webinterfaces (Django (Python), Hypernate (Java))



Primarne nukleotidne podatkovne baze

• Nucleotide sequence databases



- sequences submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents.
- entries in the EMBL, GenBank and DDBJ databases are **synchronized** on a daily basis
- **accession numbers** are managed in a consistent manner
- comparatively little error checking and fair amount of redundancy.

Nucleic sequence databases: GenBank, EMBL, and DDBJ

Genbank (NCBI - USA)
http://www.ncbi.nlm.nih.gov/Genbank/

National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search [All Databases] for [] Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More about NCBI...](#)

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources

dbGaP: NCBI's Genome Wide Association Database
NCBI's [dbGaP](#) (database of Genotype and Phenotype) provides data from Genome Wide Association (GWA) studies, which are helping elucidate the link between genes and disease. For each study, users have access to detailed information about the phenotypic variables measured and

The EMBL Nucleotide Sequence Database (EBI - UK)
http://www.ebi.ac.uk/embl/

EMBL-EBI EB-eye Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

EMBL-Bank Home
Access
Documentation
News
Submission
Publications
People
Contact

EMBL Fetch
Fetch an EMBL record by id
[] Go

IMPORTANT INFORMATION REGARDING SEQUENCE SUBMISSIONS

Fetch an EMBL record by id
[] Go

submissions ...more

Collaborations

- INSDC - International Nucleotide Sequence Database Collaboration
- NCBI - The Nucleotide Sequence Database is part of the

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 [direct submissions](#) from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international [collaboration](#) with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The [current database release](#) (Release 96, September 2008), with according [Release notes](#) and [user manual](#) are available from the EBI servers. A sample database entry is shown [here](#).

A publication in [Nucleic Acids Research](#) 2008 Oct 31. [Epub ahead of print] provides further information and details.

The EMBL nucleotide sequence database is part of the [The Protein and Nucleotide Database Group \(PANDA\)](#). This is jointly headed by [Dr. Rolf Apweiler](#) and [Dr. Ewan Birney](#), with Dr. Birney taking responsibility for Nucleotides.

Link	Explanation
Access	Database queries, Completed genomes webserver, FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA), Browse by geography.
Submission	Primary sequence submissions, third party annotation, updates.
Documentation	Release notes, user manual, information for Submitters, FAQ, Release information, Forthcoming Changes, EMBL database statistics, Feature table, XML documentation, Sample entry, Accession Number Prefix Codes, Examples of annotation, EMBL Features & Qualifiers, OE line standards, Database Policies
Publications	Group publications
People	Group members
Contact	How to contact the EMBL Nucleotide Sequence Database
News	List of recent changes on this site

DDBJ - DNA Data Bank of Japan
http://www.ddbj.nig.ac.jp/

Accession DNA Protein Taxonomy Site Search
Accession numbers
DDBJ UniProt PDB OAD PRF Patent >more

HOME Submission How to Use Search/Analysis FTP/WebAPI Report/Statistics Contact Us Japanese

About DDBJ

How to Use

Q and A

Sequence Submission

- SAKURA
- Mass Submission
- Data Updates

Search

- getentry
- ARSA
- SRS
- TXSearch
- BLAST
- PSI-BLAST
- FASTA
- SSEARCH

Phylogenetics

- ClustalW

Genome Analysis

DDBJ : DNA Data Bank of Japan
DDBJ has collaborated with EMBL/EBI and GenBank/NCBI for more than two decades to foster an archive of nucleotide sequences and their biological annotation. Namely, DDBJ is one of three summits.

Hot Topics

- Nov 25, 2008 [Release of new genome sequence data of an endosymbiont within protist cells in termite gut. 5 entries. NEW](#)
- Oct 29, 2008 [New function is added to ARSA](#)
- Oct 24, 2008 [Update of databases related to the H-Invitational](#)

Maintenance

- Nov 28, 2008 [Suspension of the DDBJ activity during the New Year Holidays](#)
- Nov 26, 2008 [NIG and DDBJ Network services temporary down](#)
- Aug 15, 2008 [\(Important\) Termination of providing SRS \(Sequence Retrieval System\) services](#)

Sequence Data Submission

- Submit my sequences
Orientation for the data submission

FTP/Web API

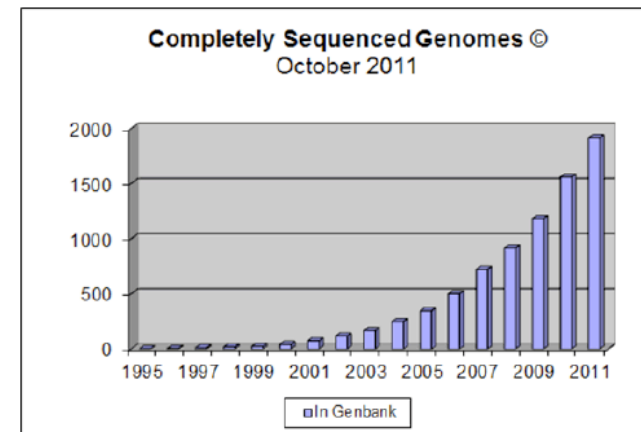
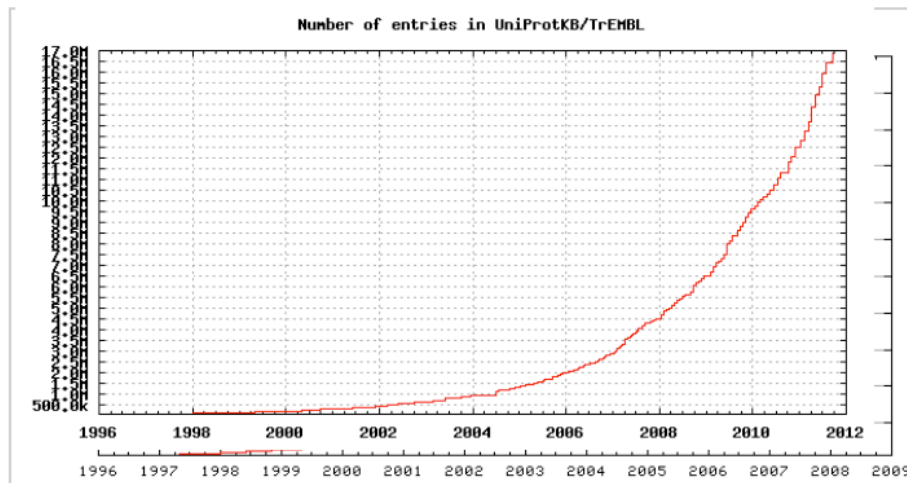
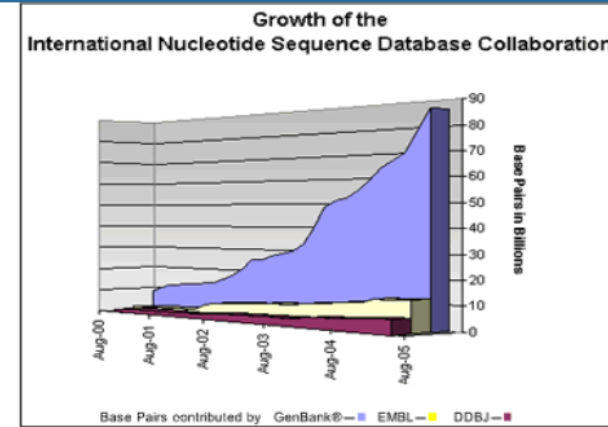
- FTP (ftp.ddbj.nig.ac.jp)
Download data files

Nucleic sequence databases

- To publish an article dealing with a sequence, scientific journals impose to have previously deposited this sequence in a reference database.
- There are 3 main repositories for nucleic acid sequences.
- Sequences deposited in any of these 3 databases are automatically synchronized in the 2 other ones.

The sequencing pace

- Nucleic sequences
 - Genbank (April 2011) <http://www.ncbi.nlm.nih.gov/genbank/>
 - 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions
 - 191,401,393,188 bases in 62,715,288 sequence records in the Whole Genome Ssequencing
- Entire genomes
 - GOLD Release V.2 (Oct 2011) contains ~2000 completely sequenced genomes.
 - http://www.genomesonline.org/gold_statistics.htm
- Protein sequences
 - Essentially obtained by translation of putative genes in nucleic sequences (almost no direct protein sequencing).
 - UniProtKB/TrEMBL (2011) contains 17 millions of protein sequences.
 - http://www.ebi.ac.uk/swissprot/sptr_stats/index.html



Adapted from Didier Gonze

Proteinske podatkovne baze

Protein sequence databases

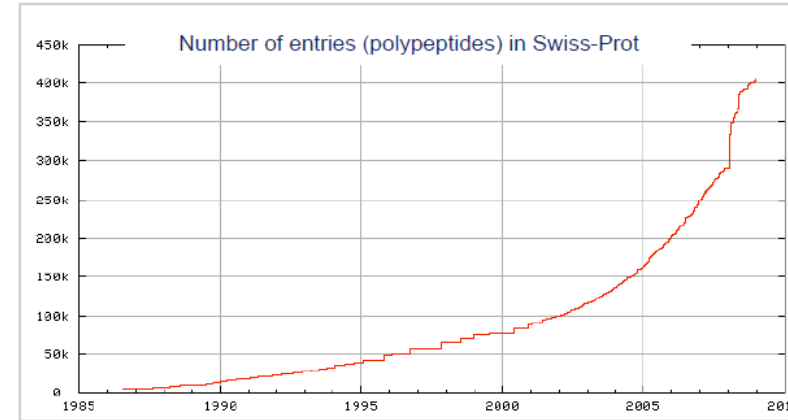


- UniProt KB
 - mission to provide a comprehensive, high-quality and freely accessible resource of protein sequence and functional information
 - **SWISS-PROT** is a protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
 - **TrEMBL** is a **computer-annotated supplement** of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.
- PIR
- SWISS-PROT and PIR are different from the nucleotide databases in that they are both **curated**

- Database content (Sept 2012)
 - UniProtKB:
 - **24,532,088 entries**
 - Translation of EMBL coding sequences (non-redundant with Swiss-Prot)
 - UniProtKB/Swiss-Prot section (reviewed):
 - **537,505 entries**
 - annotation by experts
 - high information content
 - many references to the literature
 - good reliability of the information
 - The rest (90% of the entries)
 - Automatic annotation by sequence similarity.

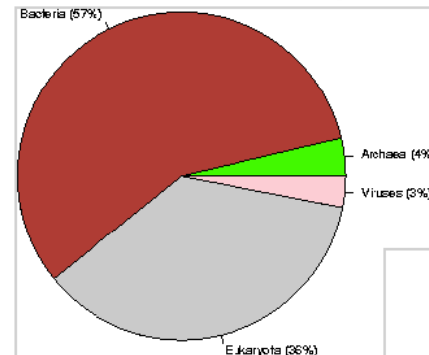
- Features
 - The most comprehensive protein database in the world.
 - A huge team: >100 annotators + developers.
 - Annotation by experts: annotators are specialized for different types of proteins or organisms.
 - World-wide recognized as an essential resource.

- References
 - Bairoch et al. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res* (1991) vol. 19 Suppl pp. 2247-9
 - The UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res* (2008). Database Issue.

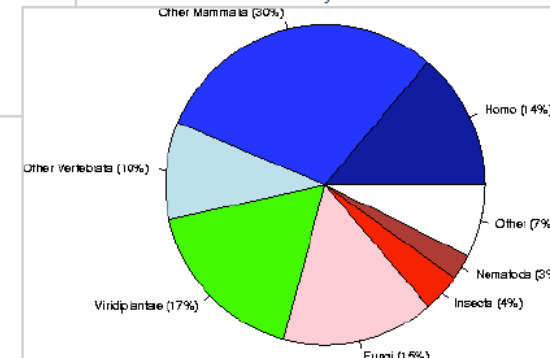


<http://www.expasy.org/sprot/relnotes/relnstat.html>

Taxonomic distribution of the sequences



Within Eukaryotes



UniProt example - Human Pax-6 protein

Header : name and synonyms

★ Reviewed, UniProtKB/Swiss-Prot **P26367** (PAX6_HUMAN) Contribute
Send feedback

Last modified November 25, 2008. Version 110. [History...](#)

Clusters with 100%, 90%, 50% identity | Documents (7) | Third-party data | Customize display TEXT XML RDF/XML GFF FASTA

[Names and origin](#) · [Protein attributes](#) · [General annotation \(Comments\)](#) · [Ontologies](#) · [Binary interactions](#) · [Alternative products](#) · [Sequence annotation \(Features\)](#) · [Sequences](#) · [References](#) · [Web resources](#) · [Cross-references](#) · [Entry information](#) · [Relevant documents](#)

Names and origin Hide | Top

Protein names	<p><i>Recommended name:</i> Paired box protein Pax-6</p> <p><i>Alternative name(s):</i> Oculorhombin Aniridia type II protein</p>
Gene names	<p>Name: PAX6</p> <p>Synonyms: AN2</p>
Organism	Homo sapiens (Human)
Taxonomic identifier	9606 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorhini › Catarrhini › Hominidae › Homo

Protein attributes Hide | Top

Sequence length	422 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is not processed.
Protein existence	Evidence at protein level.

UniProt example - Human Pax-6 protein

Human-based annotation by specialists

General annotation (Comments) Hide | Top

Function	Transcription factor with important functions in the development of the eye, nose, central nervous system and pancreas. Required for the differentiation of pancreatic islet alpha cells [By similarity] . Competes with PAX4 in binding to a common element in the glucagon, insulin and somatostatin promoters. Regulates specification of the ventral neuron subtypes by establishing the correct progenitor domains [By similarity] . Isoform 5a appears to function as a molecular switch that specifies target genes.
Subcellular location	Nucleus.
Tissue specificity	Fetal eye, brain, spinal cord and olfactory epithelium. Isoform 5a is less abundant than the PAX6 shorter form.
Developmental stage	Expressed in the developing eye and brain.
Involvement in disease	<p>Defects in PAX6 are the cause of aniridia type II (AN2) [MIM:106210]. AN2 is a bilateral panocular disorder characterized by complete or partial absence of the iris, absence of the fovea and malformations of the lens and anterior chamber. Severe age-related corneal degeneration is a frequent complication which contributes to a poor visual prognosis in aniridia. About one third of the cases are sporadic, and two thirds are familial, with autosomal dominant inheritance and high penetrance. Nearly one third of sporadic AN patients develop Wilms tumor in association with genitourinary anomalies and mental retardation (WAGR syndrome) as a consequence of heterozygous (sub)microscopic deletions of chromosome 11p13.</p> <p>Defects in PAX6 are a cause of Peters anomaly [MIM:604229]. Peters anomaly consists of a central corneal leukoma, absence of the posterior corneal stroma and Descemet membrane, and a variable degree of iris and lenticular attachments to the central aspect of the posterior cornea.</p> <p>Defects in PAX6 are a cause of ectopia pupillae [MIM:129750]. It is a congenital eye malformation in which the pupils are displaced from their normal central position.</p> <p>Defects in PAX6 are a cause of foveal hypoplasia [MIM:136520]. Foveal hypoplasia can be isolated or associated with presenile cataract. Inheritance is autosomal dominant.</p> <p>Defects in PAX6 are a cause of autosomal dominant keratitis [MIM:148190]. It is an eye disorder characterized by corneal opacification and vascularization, and by foveal hypoplasia.</p> <p>Defects in PAX6 are a cause of ocular coloboma [MIM:120200]; also known as uveoretinal coloboma or coloboma of iris, choroid and retina. Ocular colobomas are a set of malformations resulting from abnormal morphogenesis of the optic cup and stalk, and the fusion of the fetal fissure (optic fissure). Severe colobomatous malformations may cause as much as 10% of the childhood blindness. The clinical presentation of ocular coloboma is variable. Some individuals may present with minimal defects in the anterior iris leaf without other ocular defects. More complex malformations create a combination of iris, uveoretinal and/or optic nerve defects without or with microphthalmia or even anophthalmia.</p> <p>Defects in PAX6 are a cause of coloboma of optic nerve [MIM:120430].</p> <p>Defects in PAX6 are a cause of bilateral optic nerve hypoplasia [MIM:165550]; also known as bilateral optic nerve aplasia. Inheritance is autosomal dominant.</p>
Sequence similarities	<p>Belongs to the paired homeobox family.</p> <p>Contains 1 homeobox DNA-binding domain.</p> <p>Contains 1 paired domain.</p>

UniProt example - Human Pax-6 protein

Structured annotation : keywords and Gene Ontology terms

Ontologies	
Keywords	
Biological process	Differentiation Transcription Transcription regulation
Cellular component	Nucleus
Coding sequence diversity	Alternative splicing
Disease	Disease mutation
Domain	Homeobox Paired box
Ligand	DNA-binding
Molecular function	Developmental protein Repressor
Technical term	3D-structure
Gene Ontology (GO)	
Biological process	cell differentiation Inferred from electronic annotation. Source: UniProtKB-KW central nervous system development Traceable author statement. Source: Protinc eye development Traceable author statement. Source: Protinc organ morphogenesis [Ref.19] Traceable author statement. Source: Protinc regulation of transcription, DNA-dependent Inferred from electronic annotation. Source: InterPro visual perception [Ref.18] Traceable author statement. Source: Protinc
Cellular component	nucleus Inferred from electronic annotation. Source: InterPro
Molecular function	protein binding Inferred from physical interaction. Source: IntAct sequence-specific DNA binding Inferred from electronic annotation. Source: InterPro transcription factor activity [Ref.13] Traceable author statement. Source: Protinc

[Complete GO annotation...](#)

UniProt example - Human Pax-6 protein

Protein interactions; Alternative products

Binary Interactions				
With	Entry	#Exp.	IntAct	Notes
Dynl1	P63168	2	EBI-747278,EBI-349121	From a different organism.
HOMER3	Q9NSC5	2	EBI-747278,EBI-748420	
TRIM11	Q96F44	2	EBI-747278,EBI-851809	

Alternative products	
This entry describes 3 isoforms produced by alternative splicing . [Align] [Select]	
Isoform 1 (identifier: P26367-1)	This isoform has been chosen as the 'canonical' sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.
Isoform 5a (identifier: P26367-2)	Also known as: Pax6-5a; The sequence of this isoform differs from the canonical sequence as follows: 47-47: Q → QTHADAKVQVLDNQN
Isoform 3 (identifier: P26367-3)	Also known as: Pax6-5A,6*; The sequence of this isoform is not available.

UniProt example - Human Pax-6 protein

Detailed description of regions, variations, and secondary structure

Sequence annotation (Features)					
Feature key	Position(s)	Length	Description	Graphical view	Feature identifier
Molecule processing					
Chain	1 – 422	422	Paired box protein Pax-6		PRO_000050185
Regions					
Domain	4 – 130	127	Paired		
DNA binding	210 – 269	60	Homeobox		
Compositional bias	131 – 209	79	Gln/Gly-rich		
Compositional bias	279 – 422	144	Pro/Ser/Thr-rich		
Natural variations					
Alternative sequence	47	1	Q → QTHADAKVQVLDNQN in isoform 5a.		VSP_002366
Natural variant	17	1	N → S in AN2.		VAR_003808
Natural variant	18	1	G → W in AN2 and Peters anomaly.		VAR_003809
Natural variant	22 – 26	5	Missing in AN2; sporadic form.		VAR_008693
Experimental info					
Sequence conflict	317	1	R → L in AAA59963 and AAA59962. [Ref.1]		
Sequence conflict	369	1	Y → C in CAE45868. [Ref.4]		
Secondary structure					

UniProt example - Human Pax-6 protein

Peptidic sequence

Sequences				
Sequence	Length	Mass (Da)	Tools	
<input type="checkbox"/> Isoform 1 [UniParc]. Last modified July 15, 1999. Version 2. Checksum: C33CDD2C1B19C397	422	46,683	FASTA	<input type="text" value="Blast"/> <input type="button" value="go"/>
<pre> 10 LGGVFNVRG LPDSTRQKIV ELAHSGARPC DISRILQVSN GCVSKILGRY 70 YETGSIRPRA IGGSKPRVAT PEVVSIAQY KRECPISFAW EIRDRLSEGE VCTNDNIPSV 130 SSINRVLRLN ASEKQMGAD GHYDKLRMLN QOTGSWTRP GWYPQTSVPG QPTODOCQQQ 190 EGGENTNSI SBNGEDSDEA QMRLQKRKL QRNRSTFQTE QIEALEKEFE RTHYPDVFAR 250 ERLAAKIDLP EARIQVWFNS RRAKWRREEK LRNQRQASN TPSHIPISSE FSTSVYQPIP 310 QPTTPVSSFT SGMGLRGTDT ALTNYSALP PMPSPFMANN LPMQPPVPSQ TSSYSCHLPT 370 SPSVNGRSDY TYTPPHMQTH MNSQPMGTSG TTSTGLISPG VSVPPQVPDGS EFDMSQYWR EQ </pre>				
<input type="checkbox"/> Isoform 5a (Pax6-5a) [UniParc]. Checksum: 74926827347A20B5 Show >				
<input type="checkbox"/> Isoform 3 (Pax6-5A,6*) (Sequence not available).				

UniProt example - Human Pax-6 protein

References to original publications

References		Hide Top
« Hide 'large scale' references		
[1]	<p>"Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia region." Ton C.C.T., Hirvonen H., Miwa H., Weil M.M., Monaghan P., Jordan T., van Heyningen V., Hastie N.D., Meijers-Heijboer H., Drechsler M., Royer-Pokora B., Collins F.S., Swaroop A., Strong L.C., Saunders G.F. <i>Cell</i> 67:1059-1074(1991) [PubMed: 1684738] [Abstract] Cited for: NUCLEOTIDE SEQUENCE [MRNA].</p>	
[2]	<p>"Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene." Glaser T., Walton D.S., Maas R.L. <i>Nat. Genet.</i> 2:232-239(1992) [PubMed: 1345175] [Abstract] Cited for: NUCLEOTIDE SEQUENCE [MRNA].</p>	
[3]	<p>Liu J., Zhang B., Zhou Y., Peng X., Yuan J., Qiang B. Submitted (JUL-2001) to the EMBL/GenBank/DBJ databases Cited for: NUCLEOTIDE SEQUENCE (ISOFORM PAX6).</p>	
[4]	<p>The German cDNA consortium Submitted (AUG-2003) to the EMBL/GenBank/DBJ databases Cited for: NUCLEOTIDE SEQUENCE (LARGE SCALE MRNA) (ISOFORM 5A).</p>	

[24]	<p>"A novel PAX6 gene mutation (P118R) in a family with congenital nystagmus associated with a variant form of aniridia." Sonoda S., Isashiki Y., Tabata Y., Kimura K., Kakiuchi T., Ohba N. <i>Graefes Arch. Clin. Exp. Ophthalmol.</i> 238:552-558(2000) [PubMed: 10955655] [Abstract] Cited for: VARIANT NYSTAGMUS ARG-118.</p>
[25]	<p>"Missense mutation at the C-terminus of PAX6 negatively modulates homeodomain function." Singh S., Chao L.-Y., Mishra R., Davies J., Saunders G.F. <i>Hum. Mol. Genet.</i> 10:911-918(2001) [PubMed: 11309364] [Abstract] Cited for: VARIANTS AN2 GLN-375 AND ARG-422.</p>
[26]	<p>"Mutations of the PAX6 gene detected in patients with a variety of optic-nerve malformations." Azuma N., Yamaguchi Y., Handa H., Tadokoro K., Asaka A., Kawase E., Yamada M. <i>Am. J. Hum. Genet.</i> 72:1565-1570(2003) [PubMed: 12721955] [Abstract] Cited for: VARIANT MORNING GLORY DISK ANOMALY SER-68, VARIANT OCULAR COLOBOMA SER-258, VARIANT PETERS ANOMALY PRO-363, VARIANTS OPTIC NERVE HYPOPLASIA/APLASIA ILE-292; ARG-378; VAL-381 AND ALA-391.</p>
+	<p>Additional computationally mapped references.</p>

UniProt example - Human Pax-6 protein

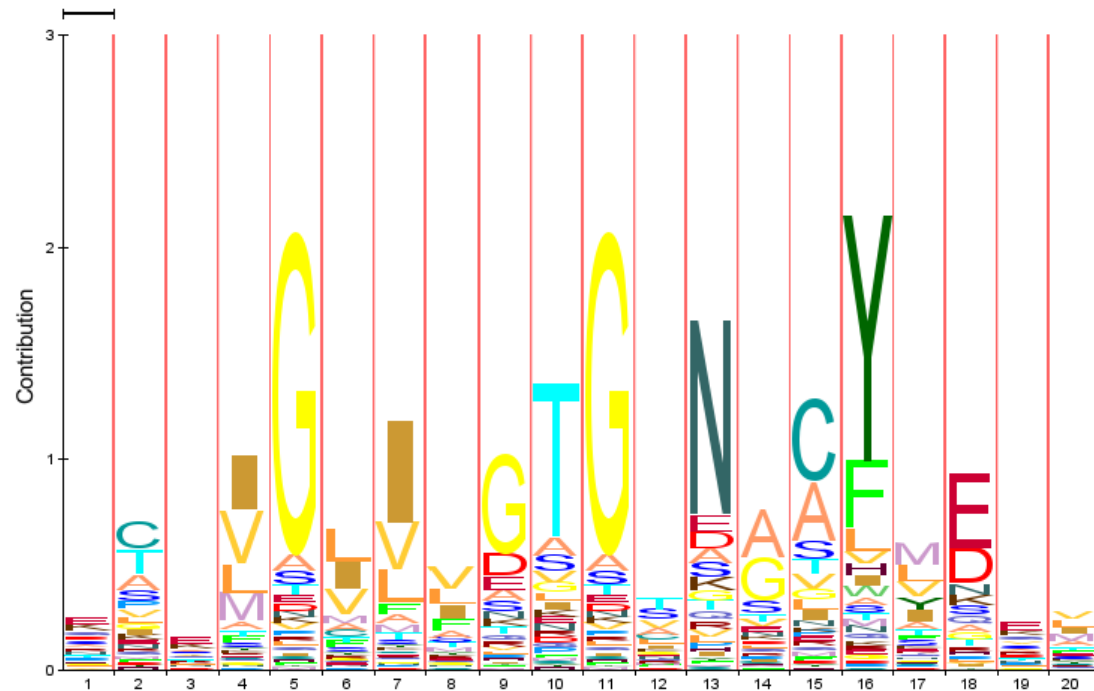
Cross-references to many databases (fragment shown)

Sequence databases																			
EMBL	M77844 mRNA. Translation: AAA59963.1. M77844 mRNA. Translation: AAA59962.1. M93650 mRNA. Translation: AAA36416.1. AY047583 mRNA. Translation: AAK95849.1. BX640762 mRNA. Translation: CAE45868.1. Z95332, Z83307 Genomic DNA. Translation: CAG38363.1. Z83307, Z95332 Genomic DNA. Translation: CAG38087.1. BC011953 mRNA. Translation: AAH11953.1.																		
PIR	A56674.																		
RefSeq	NP_000271.1. NP_001121084.1. NP_001595.2.																		
UniGene	Hs.591993																		
3D structure databases																			
PDB	<table border="1"> <thead> <tr> <th>Entry</th> <th>Method</th> <th>Resolution (Å)</th> <th>Chain</th> <th>Positions</th> <th>PDBsum</th> </tr> </thead> <tbody> <tr> <td>2CUE</td> <td>NMR</td> <td></td> <td>A</td> <td>211-277</td> <td>[*]</td> </tr> <tr> <td>6PAX</td> <td>X-ray</td> <td>2.50</td> <td>A</td> <td>4-136</td> <td>[*]</td> </tr> </tbody> </table>	Entry	Method	Resolution (Å)	Chain	Positions	PDBsum	2CUE	NMR		A	211-277	[*]	6PAX	X-ray	2.50	A	4-136	[*]
Entry	Method	Resolution (Å)	Chain	Positions	PDBsum														
2CUE	NMR		A	211-277	[*]														
6PAX	X-ray	2.50	A	4-136	[*]														
ModBase	Search...																		
Protein-protein interaction databases																			
IntAct	P26367.																		
PTM databases																			
PhosphoSite	P26367.																		
Genome annotation databases																			
Ensembl	ENSG00000007372. Homo sapiens. [Contig view]																		
GeneID	5080.																		
KEGG	hsa:5080.																		

Sekundarne biološke podatkovne baze

Secondary Databases

- Sometimes known as **pattern databases**
- Contain results from the **analysis of the sequences** in the primary databases
- Examples
 - PROSITE
 - Pfam
 - PRINTS

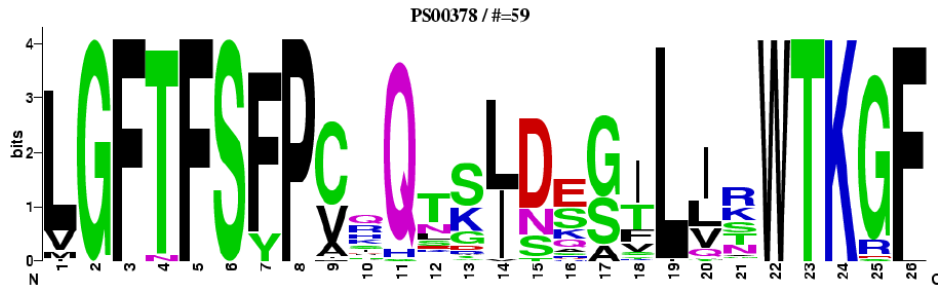


Motifs and secondary structure



PROSITE [HEXOKINASES PS00378]

- Database of protein domains, families and functional sites
- Hexokinases signature: Pattern [LIVM]-G-F-[TN]-F-S-[FY]-P-x(5)-[LIVM]-[DNST]-x(3)-[LIVM]-x(2)-W-T-K-x- [LF].

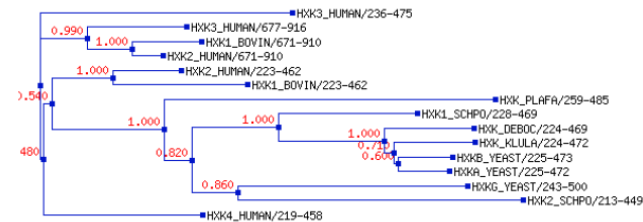


Motifs and secondary structure

Pfam [Hexokinase_2 PF03727]



- The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs)



HXK2_HUMAN

This is the summary of UniProt entry [HXK2_HUMAN](#) (P52789).

Description:	Hexokinase-2 EC=2.7.1.1
Source organism:	Homo sapiens (Human) (NCBI taxonomy View Pfam proteome data.)
Length:	917 amino acids

Please note: when we start each new Pfam data release, we take a copy of the UniProt removed after a Pfam release, these entries will not be removed from Pfam until the ne

Pfam domains

This image shows the arrangement of the Pfam domains that we found boundaries for each of the domains.



Source	Domain	Start	End
Pfam A	Hexokinase_1	16	221
Pfam A	Hexokinase_2	223	462
Pfam A	Hexokinase_1	464	669
Pfam A	Hexokinase_2	671	910



ExPASy Home page | Site Map | Search ExPASy | Contact us | Swiss-Prot | ENZYME

Search PROSITE for [Go] [Clear]

Home | ScanProsite | ProRule | Documents | Downloads | Links | Funding

Database of protein domains, families and functional sites

PROSITE consists of [documentation entries](#) describing protein domains, families and functional sites as well as associated [patterns](#) and [profiles](#) to identify them [[More details](#) / [References](#) / [Disclaimer](#) / [Commercial users](#)]. PROSITE is complemented by [ProRule](#), a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [[More details](#)].

Release 20.40, of 26-Nov-2008 (1539 documentation entries, 1315 patterns, 819 profiles and 819 ProRule)

PROSITE access

e.g: PDOC00022, PS50089, SH3, zinc

finger add wildcard ***

Browse:

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hit

SRS - Sequence Retrieval System

PROSITE tools

Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)

Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [[help](#)]:

exclude patterns with a high probability of occurrence

Prosite - aligned sequences and logo

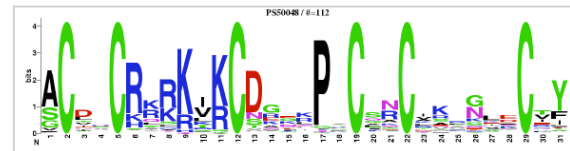
http://www.expasy.ch/prosite/

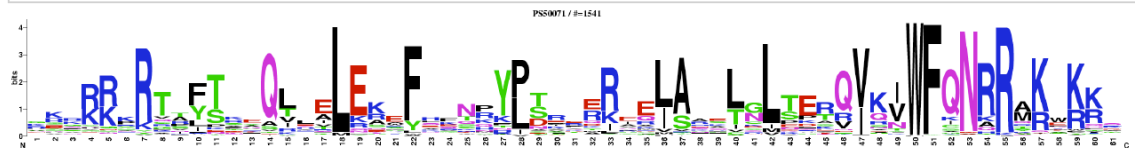
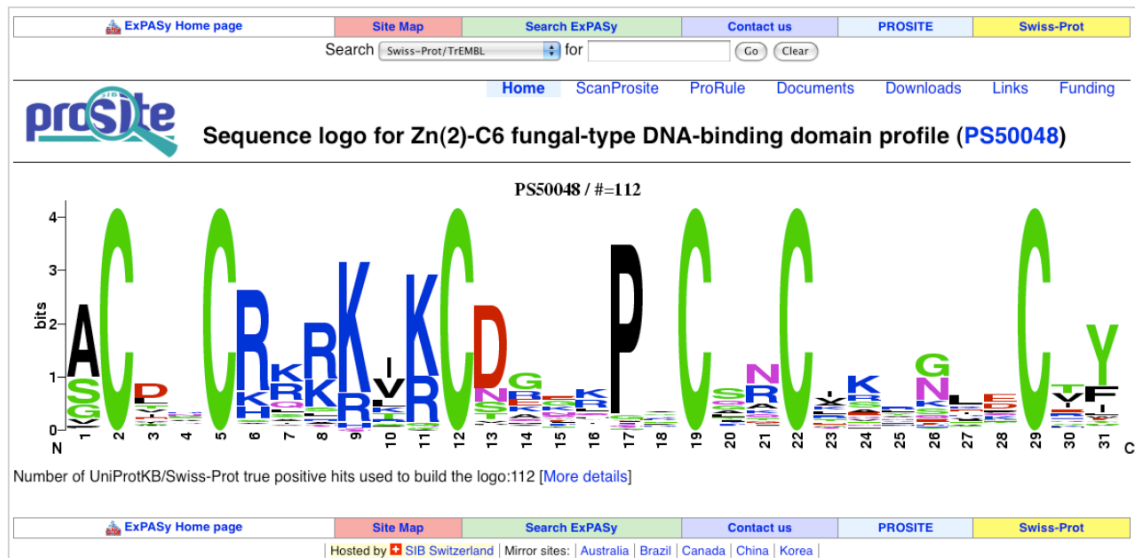


- Some of the sequences that were used to build the Prosite profile for the Zn(2)-C6 fungal-type DNA-binding domain (ZN2_CY6_FUNGAL_2, PS50048).
- The Sequence Logo (below) indicates the level of conservation of each residue in each column of the alignment.
- Note the 6 cysteines, characteristic of this domain.

```

ACE2_TREIRE/6-38      ACDRCHDKLCPRI.sGS.....CDSRQAKANV...ACVF
ACR2_NEUCR/21-49     ACYNCHRRKLCIDG...SL.....ACLKSSINGE...EC--
ACU15_NEUCR/23-53    ACDRGRSKICIDG...IR.....CDSQANVGF...ECKT
AFLR_ASPFL/28-58     SCSSCASSVICTK...EK.....ACARIERGL...ACQY
AFLR_ASPPA/28-58     SCSSCASSVICTK...EK.....ACARIERGL...ACQY
AFLR_EMENI/27-57     SCSSCASSVICTK...EK.....ACARIERGL...ACQY
ALCR_EMENI/11-51     SCSSCASSVICTK...EK.....ACARIERGL...ACQY
SCDPGRKGRCADap.eNRNeanengvsvSCSNKRWNK...DCTF
AMDR_ASPFU/26-59     ACVHCHRRVCDarivGL.....CSNCRSSGKT...DORI
AMDR_ASPOR/25-58     ACVHCHRRVCDarivGL.....CSNCRSAGKa...DORI
AMDR_EMENI/19-52     ACVHCHRRVCDarivGL.....CSNCRSAGKT...DQOI
ARGR2_YEAST/20-48    GCHTGRGRVCDL...RH.....HGORCKSNL...PC--
ARO80_YEAST/24-60    ACISGRSRVCDLgvpDNhd...pPCARCKRELK...KCIF
ATG2_PICPA/631-660   GCHTGRKROVCDL...RK.....FCLNCKESEQ...KCT-
CAT8_YEAST/69-99     ACDRGRSKTCDG...KR.....QCSQAAVGF...EERI
CBF3B_YEAST/13-44    PCSVTRRRVCDR...MI.....CGNCRKRGQd.sECHK
CHA4_YEAST/43-72     ACNCRRRRRCNM...EK.....CSNCKFRRT...EYVF
CTF1A_FUSSO/60-92    ACSTCHARRVCDAsLGV.....CTNVAFQI...EERI
CTF1B_FUSSO/52-83    ACVSGRARRVCDVv.eGA.....CGNCRWNV...EYVV
CZF1_CANAL/317-347   GCHTGRQRKICCE...TR.....KTECTRLRL...NCFW
DAL81_YEAST/149-181  SCNCRLLKTCNYf.pDLG.....NLECESTRT...KCFY
ECM22_YEAST/43-73    GCHTGRRRVCDL...GK.....FCKCTNMKL...DQVY
FCR1_CANAL/25-54     ACDSGRITKTCDG...KK.....GRRCTLDNK...IQVF
FLUF_NEUCR/10-39     ACVGRKKTCDG...QM.....GRRSRRGE...EYAY
GRL4_YEAST/10-40     ACVGRLLKLSK...EK.....KCKCKRNH...EYAY
GRT1_SCHPO/13-42     ACNCRRRVCGS...GDV.....GPECKYNE...NCFY
HAL9_YEAST/135-168   ACHGRKRICDEv...DQD...kKSNIKFQL...PQVF
HAP1_YEAST/63-95     SCTHGRKRVCCK...LR.....HGOOFTKVA.hIQY
LAC9_KLULA/94-124    ADAGRKKVCSK...TV.....TCTNLYKNL...DQVY
LEUR_YEAST/36-69     ACVEGQO.SCDah.eRA...e...PCKAKKNV...PCLL
LYS14_YEAST/158-188  GCSEKRRM.CDE...TK...TQWQARLNR...QYVY
MAL13_YEAST/12-41    ACDGRIRVCDG...KR.....CSSLQNSL...DQVY
MAL33_YEAST/7-36     ADYGRVRVCDG...KK.....CSRTEHNF...DQVY
MAL63_YEAST/7-36     SCDCGRVRVCDR...NK.....CNRCTQRNL...NCFY
MOC3_SCHPO/35-63    GCHTGRRRICDE...TK.....FCLNCTKTR...E--
NIRA_EMENI/41-72     ACTAARRRSCDG...NL.....SCAASSVYHT...TQVY
NIT4_NEUCR/52-83     NIT4ARRRSCDG...AL.....SCAASVYGT...EYIY
OAF1_YEAST/65-95     VCAQWKS.TCDR...EK.....EGGRVHKL...KQVY
PDR1_YEAST/45-74     ADRNCRKRICNG...KF.....CASCEIYSC...EYCF
PDR3_YEAST/14-43     ACVGRKRICCTG...KY.....CTNCTSYDC...TQVF
PDR8_YEAST/30-61     SCAPGRKR.LCSQ...AR.....MCOOVIRKLP...QYVY
PIP2_YEAST/24-54     VCAQRKA.TCDO...EK.....RGRGTRKQL...FQIY
PFR1_YEAST/33-63     AKRRLKICDO...EF.....SKRRAKLEV...PQVY
PRIB_LENED/19-52     ACVCRAA.MCVGA...EDGg...CQCRKANV...QCFY
PRO1_NEUCR/54-84     GCHTGRRLKICDE...GS...MCTAKHLGL...CQY
PUT3_YEAST/33-62     ACDGRKRICCPG...GN.....CCKVTSNA...IQY
    
```





Prosite - Example of domain signature

- The domain signature is a string-based pattern representing the residues that are characteristic of a domain.

ZN2_CY6_FUNGAL_1, PS00463; Zn(2)-C6 fungal-type DNA-binding domain signature (PATTERN)

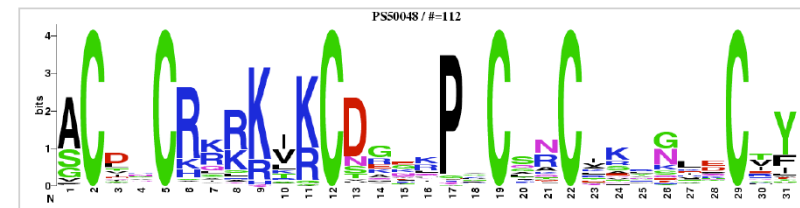
Consensus pattern: **[GASTPV] - C - x(2) - C - [RKHSTACW] - x(2) - [RKHQ] - x(2) - C - x(5,12) - C - x(2) - C - x(6,8) - C**
The 6 C's are zinc ligands

Sequences known to belong to this class detected by the pattern: ALL

Other sequence(s) detected in Swiss-Prot: human ultra high-sulfur keratin.

- Retrieve an alignment of Swiss-Prot true positive hits: [Clustal format, color, condensed view](#) / [Clustal format, color](#) / [Clustal format, plain text](#) / [Fasta format](#)
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS00463
- Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS00463
- Scan Swiss-Prot/TrEMBL entries against PS00463
- view ligand binding statistics

Matching PDB structures: 1AJY 1AW6 1CLD 1D66 ... [\[ALL\]](#)



PFAM (Sanger Institute - UK) <http://pfam.sanger.ac.uk/>
 Protein families represented by multiple sequence alignments and hidden Markov models (HMMs)

welcome trust **sanger** institute

HOME | SEARCH | BROWSE | FTP | HELP

Pfam
keyword search

Family: Zn_clus (PF00172)

81 architectures 3469 sequences 2 interactions 85 species 24 structures

Summary

Domain organisation

Alignments

HMM logo

Trees

Curation & models

Species

Interactions

Structures

Jump to...

Summary

Fungal Zn(2)-Cys(6) binuclear cluster domain

No Pfam abstract.

Interpro entry [IPR001138](#)

The N-terminal region of a number of fungal transcriptional regulatory proteins contains a Cys-rich motif that is involved in zinc-dependent binding of DNA. The region forms a binuclear Zn cluster, in which two Zn atoms are bound by six Cys residues [PUBMED:2107541](#), [PUBMED:1552122](#). A wide range of proteins are known to contain this domain. These include the proteins involved in arginine, proline, pyrimidine, quinone, maltose and galactose metabolism; amide and GABA catabolism; leucine biosynthesis and others.

Gene Ontology

Cellular component	nucleus (GO:0005634)
Molecular function	zinc ion binding (GO:0008270)
Biological process	regulation of transcription, DNA-dependent (GO:0006355)

Internal database links

SCOP: [EndIII_4Fe-2S](#)

External database links

HOMSTRAD:	GAL4
PANDIT:	PF00172
PRINTS:	PR00054
PROSITE:	PDOC00378
SCOP:	1d66
SYSTEMS:	Zn_clus

Example structure
 PDB entry [1d66](#): DNA RECOGNITION BY GAL4: STRUCTURE OF A PROTEIN-DNA COMPLEX

InterPro (EBI - UK)
<http://www.ebi.ac.uk/interpro/>

EBI > Databases > InterPro

Search InterPro:

InterPro: Home

InterPro is a database of protein families, domains, repeats and sites in which identifiable features found in known proteins can be applied to new protein sequences.

Release News

Announcement:

- InterPro 18.0 is released** and covers 75.6% of UniProtKB, with new methods from PROSITE, GENESD and SUPERFAMILY.
- PROSITE pattern matches** are now evaluated to either TRUE (T) or UNKNOWN (U) using miniprofiles or associated existing PROSITE profiles.

Please see [Release Notes](#) for further details.

General Information:

- Match complete.xml (UniProtKB) now contains all UniProtKB proteins including those not matching an InterPro signature.
- UniParc (uniparc_match.tar.gz) and UniMES (unimes_match.tar.gz) matches to InterPro methods have been updated and are available from the [ftp site](#) in XML format.

Note: due to the large size of UniParc and UniMES the data has been divided into chunks and the latest updates are provided in these files at each InterPro release.

Future proposed changes:

InterPro will be introducing new entry classification rules that will affect how an entry is typed:

- Entries typed **Repeat** or **Site** will remain the same.
- Entries typed **Family** or **Domain** will follow stricter criteria to ensure they conform more closely to current biological concepts:
 - Entries typed **Family** will contain signatures that cover all domains in the matching proteins.
 - Entries typed **Domain** will identify biological units with defined boundaries, which includes structural domains/subdomains as well as functional domains.
 - All remaining entries will be covered by a new type, **Region** including those which cover more than one domain, as well as those covering partial domains.
- New relationship rules will be introduced that will affect how different entries are related to one another. **ParentChild** and **Contains/Found in** relationships will continue within InterPro with their existing definitions, but the following changes will occur:
 - Entry type will no longer have any bearing on the relationships of that entry. Instead, only the sequence covered by the signatures of an entry will be taken into consideration when forming relationships.
 - ParentChild** relationships will be permitted between entries of different types.
 - All **Contains/Found in** relationships for an entry will be displayed in the Relationships section of an entry (currently, only the most specific are displayed).

Any concerns or comments regarding the proposed changes should be directed to [EBI Support](#).

User support and feedback

We welcome feedback, particularly if you find errors or omissions please let us know. If you need information

InterPro (EBI - UK)
 Antennapedia-like Homeobox (entry IPR001827)

EBI > Databases > InterPro

Jump to: [InterProScan](#) [Databases](#) [Documentation](#) [FTP site](#) [Help](#) [Advanced search](#)

Search InterPro:

InterPro: IPR001827 Homeobox protein, antennapedia type

Protein matches

UniProtKB Matches: 742 proteins

Overview: [sorted by AC](#), [sorted by name](#), [of known structure](#), [proteins with splice variants](#)

Detailed: [sorted by AC](#), [sorted by name](#), [of known structure](#), [proteins with splice variants](#)

Table: [For all matching proteins](#), [of known structure](#)

[Architectures](#)

[Accession List](#)

Accession [IPR001827](#) Antennapedia

Type [Domain](#)

Database	ID	Name	Proteins
PRINTS	PS00025	ANTENNAPEAIA	510
PROSITE pattern	PS00032	ANTENNAPEAIA	973

InterPro Relationships

Parent [IPR01356](#) Homeobox

GO Term annotations

Process [GO:0006355](#) regulation of transcription, DNA-dependent

Function [GO:0003677](#) DNA binding
[GO:0003700](#) transcription factor activity

InterPro annotation

The homeobox is a 60-residue motif first identified in a number of *Drosophila* homeotic and segmentation proteins, but now known to be well-conserved in many other animals, including vertebrates [1, 2, 3]. Proteins containing homeobox domains are likely to play an important role in development - most are known to be sequence-specific DNA-binding transcription factors. The domain binds DNA through a helix-turn-helix (HTH) structure.

Many homeodomain-containing proteins have now been sequenced and, while the homeodomain flanking regions vary, characteristic conserved sequences upstream of the domain allow the proteins to be grouped into 3 subfamilies: the so-called antennapedia, engrailed and paired box proteins. Antennapedia, which regulates the formation of leg structures in *Drosophila*, was one of the first homeotic genes studied and led to the discovery of the homeobox domain. Over expression of this gene in the wrong segment of the fruit fly can lead to the formation of leg structures in these segments. For example, over expression in the head segment can lead to the formation of legs instead of antennae (hence the name antennapedia). The sequences of the antennapedia proteins contain a conserved hexapeptide 5-16 residues upstream of the homeobox, the specific function of which is unclear. The six *Drosophila* proteins that belong to this group are antennapedia (Antp), abdominal-A (abd-A), deformed (Dfd), proboscipedia (pb), sex combs reduced (scr) and ultrabithorax (ubx) and are collectively known as the 'antennapedia' subfamily.

In vertebrates the corresponding Hox genes are known [4] as Hox-A2, A3, A4, A5, A6, A7, Hox-B1, B2, B3, B4, B5, B6, B7, B8, Hox-C4, C5, C6, C8, Hox-D1, D3, D4 and D6.

Caenorhabditis elegans lin-39 and mab-5 are also members of the 'antennapedia' subfamily.

Arg and Lys are most frequently found in the last position of the hexapeptide; other amino acids are found in only a few cases.

Structural links

PDB: [click here](#)

SCOP: [s.d.1.1](#)

CATH: [1.10.10.60.20](#), [1.10.10.60.4](#)

Database links

MSDsite: [PS00032](#)

PROSITE doc: [PDOC00032](#)

Blocks: [IPR001827](#)

Peptide related information

- [MEROPS](#) - Peptidase Database
- [Peptide Database \(Cancer\)](#) [[example](#)]
- [PeptideMass](#)
 - cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot and TrEMBL) or a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides.
- [SYFPEITHI](#)
 - SYFPEITHI is a database comprising more than 7000 peptide sequences known to bind class I and class II MHC molecules. The entries are compiled from published reports only.
- [PeptideAtlas](#)
 - multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments.



Composite or Meta databases

These databases of databases collect data from different sources and make them available in new and more convenient form, or with an emphasis on a particular disease or organism.

Entrez – main page

Free text search

Pick a database

NCBI
National Center for Biotechnology Information
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMM Books TaxBrowser Reference

Search All Databases for [] Go

SITE MAP
Alphabetical Resource
About NCBI
GenBank
Literature databases
Molecular databases
Genomics

What does NCBI do?
Hot Spots

igabases
The new My NCBI has replaced the old Entrez search interface and includes automatic search updates and filtering search results.

Done Internet

Many Datasets at NCBI

- The NCBI hosts a huge interconnected database system that, in addition to DNA and protein, includes:
 - Journal Articles (PubMed)
 - Genetic Diseases (OMIM)
 - Polymorphisms (dbSNP)
 - Cytogenetics (CGH/SKY/FISH & CGAP)
 - Gene Expression (GEO)
 - Taxonomy
 - Chemistry (PubChem)



Home Analyze Sequence Function Literature Community

New & Noteworthy
Seminal Yeast Literature
 08/27/2013
 SGD has compiled a selection of seminal yeast literature, comprising landmark papers in yeast biology. The list is available on the SGD Wiki and includes important publications on cell biology, early genetic maps and genome surveys, and the original S288C sequencing consortium. Also listed are key papers describing strains of *S. cerevisiae*. This new already available on the SGD...re

About SGD
 The *Saccharomyces* Genome Database (SGD) provides comprehensive integrated biological information for the budding yeast *Saccharomyces cerevisiae* along with search and analysis tools to explore these data, enabling the discovery of functional relationships between sequence and

FB2013_04, released July 15th, 2013

FlyBase
 A Database of *Drosophila* Genes & Genomes

Home Tools Files Species Documents Resources News Help Archives Jump to Gene Go

BLAST GBrowse QueryBuilder RNA-Seq Search TermLink ImageBrowse Batch Download

Fast-Track Your Paper
 FlyBase Forum
 Find a Fly Person

QuickSearch
 Simple Expression Phenotype GO References Data Type

Species: include non-Dmel species

Enter text:

Note: Wild cards (*) can be added to your search term

WormBase Version: WR258

Species Resources Tools

Explore Worm Biology
 facilitating insights into nematode biology

News
 Changes at WormBase
 WormBase thanks Jonathan Haldin for years of superb community service handling genetic nomenclature. The School will now serve as the genetic nomenclature coordinator and advisor. [Join Sarah, Benjamin Beer and The School](#)

Activity
 Random page:
Caenorhabditis elegans EST, clone B01, end frame 003562, GBrowse_3_371022P
 371022.paf: This EST sequence belongs to 1 or more Mapped/Cloned data submissions

What's popular on WormBase:
[Burs on history](#)

You can save items on WormBase!
 When you see a star on WormBase!

Commentary See all commentaries

FlyBase RNA-Seq RPKM data bulk download

digestive system, 1-day adult	2
digestive system, 4-day adult	4
digestive system, 20-day adult	1
fat body, larval L3 wandering	4
fat body, white prepupae	40
fat body, pupal PB	25
cornuata, larval L3 wandering	15
cornuata, 1-day adult	15
cornuata, 4-day adult	95
cornuata, 20-day adult	11
ovary, single 4-day female	66
ovary, mated 4-day female	24
ovary, mated 4-day male	65

May 6, 2013. FlyBase is extending its initial gene-level analyses of RNA-seq throughput data from modENCODE and others. The algorithm for RPKM (reads per kilobase per million mapped reads) has been refined, additional datasets have been analyzed, and these data are now available for bulk download... (More)

NCBI - Welcome Website

NCBI - Search All Databases

NCBI - Search All Databases

Database: {All Databases}

Search term / query: Daphnia

NCBI - Search Across Databases - Summary

NCBI - Search Literature Database - PubMed Central (free)

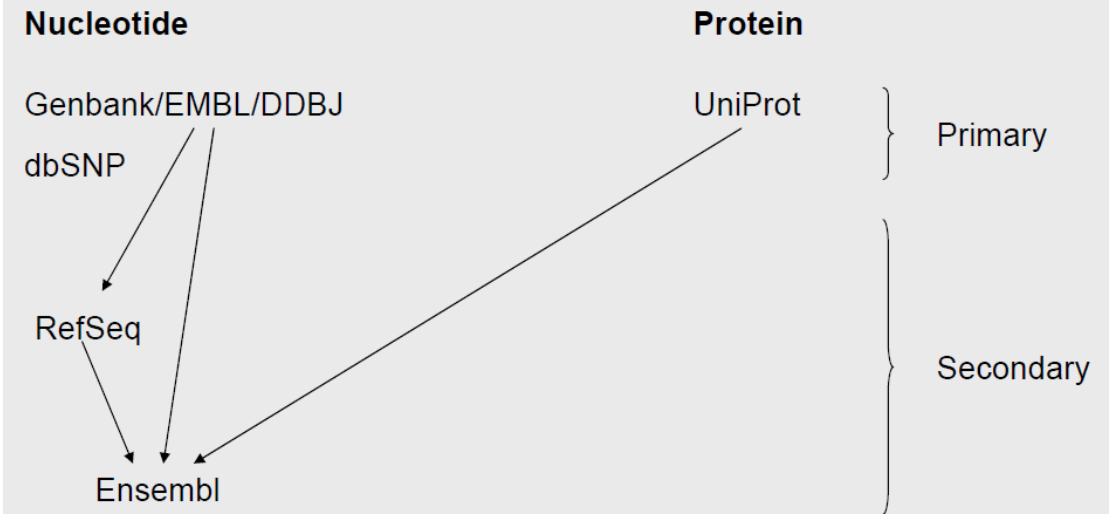
NCBI - PubMed Central (PMC)

Integrating databases

Why integration?

- **Data is distributed to several sources**
 - That can prevent efficient access to data
- **Genomics**
 - Study of whole genomes, knowledge of gene content, expression etc. needed
- **To get a better view to cells**
 - Systems biology
 - Reductionism doesn't work by itself anymore, we need integration of knowledge
 - One PhD student, one gene ;(
 - Add protein studies, metabolomics, etc.

Hierarchy of databases - an illustrative example



Data integration

- **Biomart** www.biomart.org


BioMart is a query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI).

The system can be used with any type of data and is particularly suited for providing 'data mining' like searches of complex descriptive data.

- **Web services** www.biocatalogue.org

Web services are application programming interfaces (API) or web APIs that are accessed via Hypertext Transfer Protocol and executed on a remote system hosting the requested services

Specializirane biološke podatkovne baze



the Gene Ontology

gene or protein name

Open menus

Home

FAQ

Downloads

Tools

Documentation

About GO

Projects

Contact GO

Site Map

Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. [Read more about the Gene Ontology...](#)

Search the Gene Ontology Database

Search for genes, proteins or GO terms using [AmiGO](#):

gene or protein name
 GO term or ID

AmiGO is the official GO browser and search engine. [Browse the Gene Ontology with AmiGO.](#)

GO website

- The latest news and views in the [GO newsletter](#)
- GO downloads, including [ontology files](#), [annotations](#) and the [GO database](#)
- Tools for using GO, including [OBO-Edit downloads](#), [AmiGO](#), and the [GO Online SQL Environment](#).
- [Request new terms or ontology changes](#) or [get help with new term submission](#)
- Documentation on all aspects of the GO project and the [GO FAQ](#)
- Projects within the GO consortium, including [Reference Genomes](#) and [immune system annotation](#)
- [Gene Ontology mailing lists](#) and [contact details](#)

The Gene Ontology Consortium is supported by a P41 grant from the National Human Genome Research Institute (NHGRI) [grant [HG002273](#)]. See the [full list of funding sources](#). The Gene Ontology Consortium would like to acknowledge the assistance of many more people than can be listed here. Please visit the [acknowledgements page](#) for the full list.



open biomedical ontologies

Last modified Wednesday, 19-Mar-2008 17:10:11 PDT

[Cite GO](#) • [Terms of use](#) • [GO helpdesk](#)

Copyright © 1999-Saturday, 29-Nov-2008 17:49:09 PST [the Gene Ontology](#)

Gene Ontology Database (<http://www.geneontology.org/>)

Example: methionine biosynthetic process

- ☐ all : all [251524 gene products]
 - ☐ **GO:0008150** : biological_process [165760 gene products]
 - ☐ **GO:0009987** : cellular_process [78832 gene products]
 - ☐ **GO:0044237** : cellular_metabolic_process [53731 gene products]
 - ☐ **GO:0006519** : cellular_amino_acid_and_derivative_metabolic_process [4751 gene products]
 - ☐ **GO:0006520** : amino_acid_metabolic_process [3961 gene products]
 - ☐ **GO:0008652** : amino_acid_biosynthetic_process [1807 gene products]
 - ☐ **GO:0009067** : aspartate_family_amino_acid_biosynthetic_process [485 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0000097** : sulfur_amino_acid_biosynthetic_process [288 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0009066** : aspartate_family_amino_acid_metabolic_process [714 gene products]
 - ☐ **GO:0009067** : aspartate_family_amino_acid_biosynthetic_process [485 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0006555** : methionine_metabolic_process [281 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0000096** : sulfur_amino_acid_metabolic_process [446 gene products]
 - ☐ **GO:0006555** : methionine_metabolic_process [281 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0000097** : sulfur_amino_acid_biosynthetic_process [288 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0044249** : cellular_biosynthetic_process [27813 gene products]
 - ☐ **GO:0044271** : nitrogen_compound_biosynthetic_process [2165 gene products]
 - ☐ **GO:0009309** : amine_biosynthetic_process [1996 gene products]
 - ☐ **GO:0008652** : amino_acid_biosynthetic_process [1807 gene products]
 - ☐ **GO:0009067** : aspartate_family_amino_acid_biosynthetic_process [485 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0000097** : sulfur_amino_acid_biosynthetic_process [288 gene products]
 - ☐ **GO:0009086** : **methionine biosynthetic process** [171 gene products]
 - ☐ **GO:0044272** : sulfur_compound_biosynthetic_process [548 gene products]

Actions...

Last action: Reset the tree

[Graphical View](#)

[View in tree browser](#)

[Download...](#)

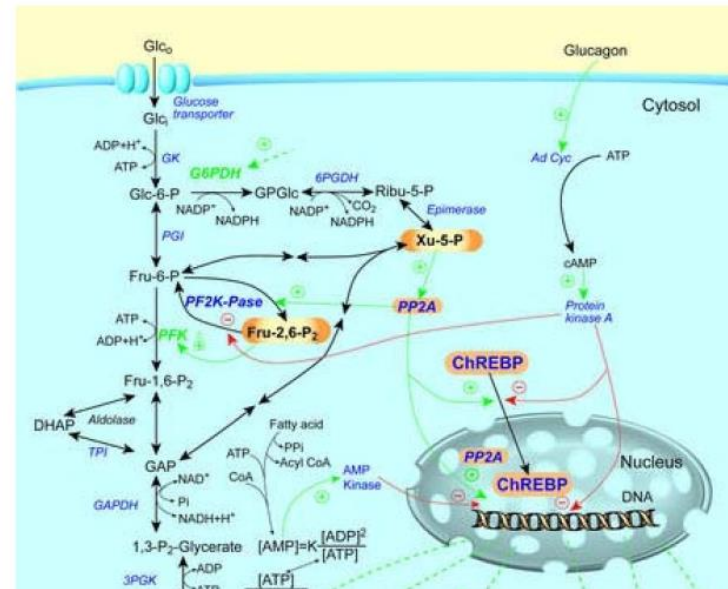
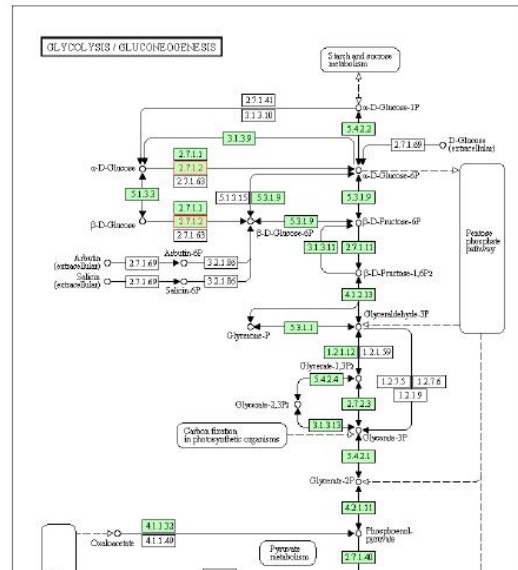
[OBO](#)

[RDF/XML](#)

[GraphViz dot](#)

Metabolic networks - pathways

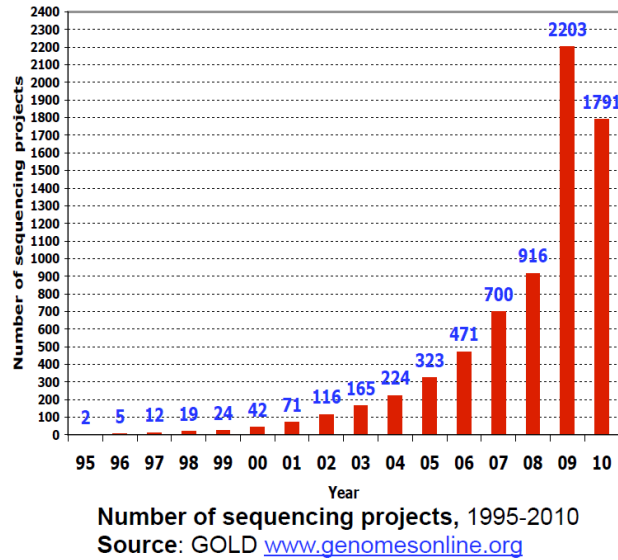
- Kegg Pathways [glycolysis / gluconeogenesis hsa]
- MetaCyc (HumanCyc)
- Reactome - a curated knowledgebase of biological pathways



Genomske podatkovne baze

Evolution of genome projects

- 7464 genome projects (June 2010)
- Number and Size of projects grows at a rapid rate
- We have crossed the Terabyte threshold of genomic data
- These data need to be stored, curated and made available for analysis and knowledge discovery



Database definition

- What is a database ?
 - An organized body of related data
 - It is presumed that (i) the volume of the data is large (ii) data have to be accessed, retrieved, updated “frequently”
 - Database Management Systems (DBMS) are software that facilitates the creation, administration and use of the database

• Some common database models:

Flat File Model

Record	Route No.	Miles	Activity
Record 1	I-95	12	Overlay
Record 2	I-495	05	Patching
Record 3	SR-301	33	Crack seal

Structured Files

Relational Model

EMPLOYEE

ID : NUMBER
F_NAME : VARCHAR
L_NAME : VARCHAR
SALARY : NUMBER
ADDRESS_ID : NUMBER

←

EMP_DATA

EMP_ID : NUMBER
MGR_ID : NUMBER
YEAR_OF_SERV : NUMBER

Tables

Object-Oriented Model

Object 1: Maintenance Report

Date: 01-12-01
Activity Code: 24
Route No: I-95
Daily Production: 2.5
Equipment Hours: 6.0
Labor Hours: 6.0

Object 2: Maintenance Activity

Activity Code:
Activity Name:
Production Unit:
Average Daily Production Rate:

Objects

How big are genomes ?

GenBank Release 173.0 — August 15, 2009

Species	Genome size	Bases	Entries
Homo sapiens	3,400,000,000	13,669,851,495	12,838,795
Mus musculus	3,454,200,000	8,445,993,792	7,347,636
Rattus norvegicus	2,900,000,000	6,284,206,670	1,997,976
Bos taurus	3,651,500,000	5,319,815,212	2,135,747
Zea mays	5,000,000,000	5,007,807,286	3,870,406
Sus scrofa	3,108,700,000	4,229,790,475	2,536,492
Danio rerio	1,900,000,000	3,074,615,557	1,695,362
Strongylocentrotus purpur	900,000,000	1,352,840,985	228,153
Nicotiana tabacum	900,000,000	1,184,330,809	1,752,654
Oryza sativa Japonica Gro	900,000,000	1,176,024,629	1,217,983
Xenopus (Silurana) tropic	900,000,000	1,146,732,476	1,423,046
Drosophila melanogaster	180,000,000	1,038,512,618	1,202,127
Pan troglodytes	3,577,500,000	997,816,950	213,217
Arabidopsis thaliana	100,000,000	950,139,115	2,240,601
Canis lupus familiaris	100,000,000	931,176,470	1,434,100
Vitis vinifera	100,000,000	910,760,908	655,658
Gallus gallus	1,200,000,000	884,489,747	806,871
Glycine max	1,115,000,000	846,429,180	1,828,912
Macaca mulatta	3,543,000,000	808,403,289	78,410
Ciona intestinalis	200,000,000	748,153,905	1,216,132
Total		106,533,156,756	108,431,692

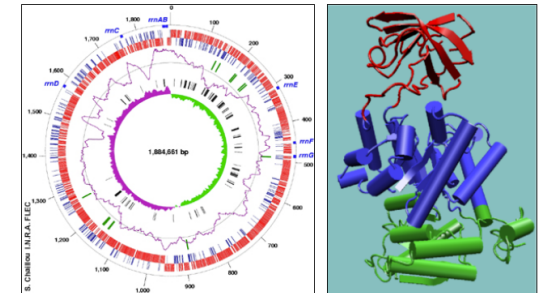
DOGS – Databases Of Genome Sizes
www.cbs.dtu.dk/databases/DOGS

Genome typical sizes:

- Virus: 1 to 360 kb (mimivirus: 1.2 Mb !)
- Bacteria: 0.5 to 13 Mb
- Eukaryotes: 8 Mb to 670 Gb

What are the genomic data ?

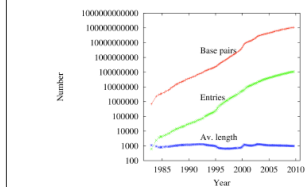
- Genomes
- Chromosomes
- Genes (cDNA, ESTs, RNAseq,..)
- Nucleotide sequences
- Protein sequences
- Annotations
- Structures
- Etc...



```

ID AB001437.2; parent: AB001437
AC AB001437; AB007513-AB007868;
FT gene             467..1807
FT                 /gene="dmaa"
FT                 /locus_tag="CA_C0001"
SQ
Sequence 1341 BP;
atggctgcc aattaatga actgtgcga aaacatata acataatata agtgaatta    60
acggaatga gcttaaacg tggataaaa agtatacctc ctatctatc tggataaagt    120
agctacgca ttagcttcc aatacaatt caaaagaaa ttctopaaa tgaataaaa    180
gattataaa taattctat gaataataa actactaaa aatagatcat agcatatta    240
atttcctctg aagaacttt gaaacgcat gaagataaa aaacagataa aaacagata    300
aatcagata ctatgcttc tatgcttat ccaataaca aattgcttc attgtaatt    360
ggataatga atagatttc tcaagtga tgcctgctg ttgctgaga tctctgaaa    420
gctacatc  cattttat atagagga gttggactg gaagactca tttatgca    480
gctatagtc attacatct tgaataat caaaagcta agtgtgata tgtttctt    540
    
```

Top 20 organisms in public banks



Center for Biological Sequence Analysis
The Technical University of Denmark

Data visualization

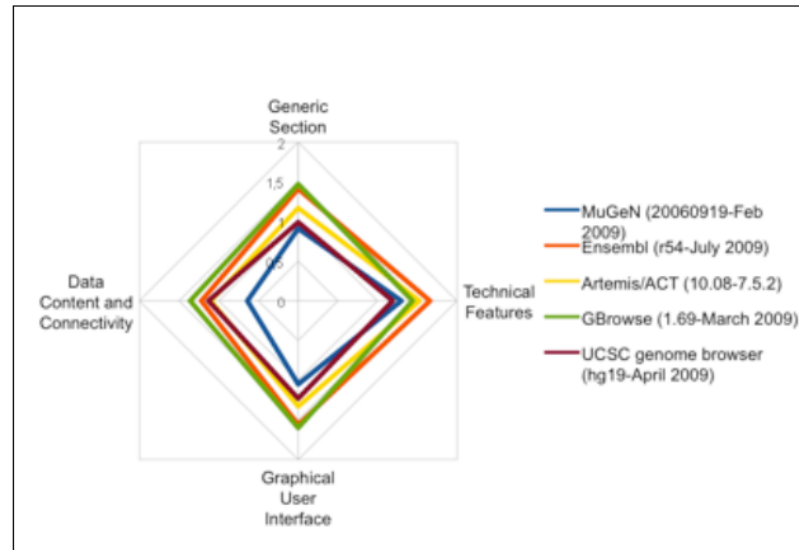
Genome Browsers (GBs) hold a central place in genomic projects

- Some popular GBs: Gbrowse, Artemis, Ensembl, UCSC,...

- Many GBs are now available but the choice of a well adapted GB can be a challenging task

- Web site:

genome.jouy.inra.fr/CompaGB



Main resources for genomes

- Two main international resources:
 - **EBI: European Bioinformatics Institute**
www.ebi.ac.uk/genomes
 - **NCBI: National Center for Biotechnology Information**
www.ncbi.nlm.nih.gov/Genomes
- Many other resources from sequencing institutions:
 - **Sanger:** the Wellcome Trust Sanger Institute www.sanger.ac.uk
 - **JCVI:** the Craig Venter Institute (formerly TIGR) www.jcvi.org
 - **JGI:** the Joint Genome Institute (DOE) genome.jgi-psf.org
 - **Broad:** the Broad Institute (Harvard, MIT) www.broadinstitute.org
 - **Genoscope:** the french sequencing center www.genoscope.cns.fr
 -

Browsing genomes at EBI

Genomes Pages - At the EBI

Access to Completed Genomes

The first completed genomes from viruses, phages and organelles were deposited into the EMBL Database in the early 1980's. Since then, molecular biology's shift to obtain the complete sequence of as many genomes as possible combined with major developments in sequencing technology resulted in hundreds of complete genome sequences being added to the database, including Archaea, Bacteria and Eukaryota. These web pages give access to a large number of complete genomes. Data is available to describe the layout.

Whole Genome Shotgun Sequences (WGS)

Methods using whole genome shotgun data are used to gain a large amount of genome coverage for an organism. WGS data for a growing number of organisms are being submitted to DDBJ/EMBL/GenBank and are made available via EBI's Sequence Retrieval System (SRS) at <http://srs.ebi.ac.uk> and the EBI FTP server at <ftp://ftp.ebi.ac.uk/pub/databases/embldb/wgs/>. More information about WGS projects...

Human Draft Genome

The completion of the human draft genome sequence was announced and published in February 2001 in *Nature* and *Science*. Since the beginning of the Human Genome Project, the International Human Genome Sequencing Consortium has been submitting human draft sequence data to the International Nucleotide Sequence Databases DDBJ/EMBL/GenBank. High-throughput human sequences have been made available to the public immediately via the EMBL Database High-throughput genome directory (HTG), while finished sequences have been included in the Human Genome Project.

Genome Annotation and Proteome Analysis

Genome Reviews are a set of reannotated proteomes, based on manually curated data from UniProt.

The EMBL Genome Browser provides the best possible automatic annotation, graphical views and web-searchable datasets for a number of eukaryotic genomes including human, mouse, zebrafish, anopheles, drosophila and others to follow.

Integr8 has proteome analysis information on a large number of organisms.

Last 40 Genome Entries

NB The genomes pages are normally updated ahead of SRS. If links fail to work for recent genomes, please try in 1-2 days.

Date	Accession	Description
01-JUN-2010	CP001958.1	Sequilaria rotundus DSM 44985
27-MAY-2010	GB845441.1	Human papillomavirus type 119
27-MAY-2010	GB845442.1	Human papillomavirus type 120
27-MAY-2010	GB845443.1	Human papillomavirus type 121
27-MAY-2010	GB845444.1	Human papillomavirus type 122
27-MAY-2010	GB845445.1	Human papillomavirus type 123

EBI: genomes of archaea

Genomes Pages - Archaea

84 organisms.

Accession numbers of all the entries listed below may be downloaded as a text file for use in downloading using the [Sequence Version Archive](#).

A more-detailed, [tab-delimited list](#) is also available.

Description	Length (bp)	Sequence		Project	Proteome
		Plain	HTML		
1 species <i>Aciduliferundum boonei</i>	1,488,778	CP001941.1	CPW01941	38403	Proteome
2 species <i>Aeropyrum pernix K1</i>	1,688,696	BA000002	BA000002	211	Proteome
3 species <i>Archaeoglobus fulgidus</i>	2,178,400	AE000782	AC000786	104	Proteome
4 species <i>Archaeoglobus profundus</i>	1,560,622	CP001857	CPW01857	32583	Proteome
5 species <i>Caldivirga maquilingsensis</i>	2,077,567	CP000852	CPW00852	17421	Proteome
6 species <i>Candidatus Korarchaeum cryptofilum</i>	1,590,757	CP000968	CPW00968	16525	Proteome
7 species <i>Candidatus Methanoregula boonei</i>	2,542,943	CP000780	CPW00780	18505	Proteome
8 species <i>Conarchaeum symbiosum</i>	2,045,086	DP000238	DPW00238	202	Proteome
9 species <i>Desulfurococcus kamchatkensis</i>	1,365,223	CP001110	CPW01110	28141	Proteome
10 species <i>Ferroplasma placidus</i>	2,196,266	CP001899	CPW01899	33635	Proteome
11 species <i>Ferroplasma acidimanus</i>	1,947,983	CM000428	CMW00428	239	n/a

Genome databases

EBI: genome of *A. pernix K1*

Integr8 - A.pernix

Integr8 - A.pernix

Search for species: Search for gene/protein: in **A.pernix**

Selected species **A.pernix** Change scope

Aeropyrum pernix (strain K1) - Tax ID: 272557 GAS: [FASTA](#)

A strictly aerobic hyperthermophilic archaeon isolated from a coarsely sulfataric thermal vent at Kodakara-Jima Island, Japan in 1993. It is a heterotroph that grows optimally at 90 to 95 degrees Celsius, pH 7.0, and a salinity of 3.5%. It is spherical shaped and covered by a cell envelope (S-layer-like structure). It contains C25-isoprenyl archaeal (glycerol diether) as the hydrocarbon chain in the core lipid.

Literature Genome Statistics Proteome Analysis Taxonomy Downloads

Integr8 - A.pernix Genome Statistics

Integr8 - A.pernix Genome Statistics

Search for species: Search for gene/protein: in **A.pernix**

Component name	Protein count	Type	Length (bp)	Avg. CDS Length	CDS constant	CDS coverage	Gene count
Chromosome	1709	C	1688888	871.376	86.3%	86%	1709

Genome Statistics: [Protein count](#) [Type](#) [Length \(bp\)](#) [Avg. CDS Length](#) [CDS constant](#) [CDS coverage](#) [Gene count](#)

Genome Statistics: [Protein count](#) [Type](#) [Length \(bp\)](#) [Avg. CDS Length](#) [CDS constant](#) [CDS coverage](#) [Gene count](#)

II. Genome databases

EBI: genomes of eukaryotes

Genomes Pages - Eukaryota

114 organisms.

Accession numbers of all the entries listed below may be downloaded as a [text file](#) for use in downloading using the [Sequence Version Archive](#). Due to the increased numbers of completed genome sequences, this page no longer includes direct links to [Ensembl genomes](#). Please use the link to browse them directly.

A more detailed, [tab-delimited list](#) is also available.

	Description	Length (bp)	Sequence		Project	Proteins
			Plain	HTML		
species <i>Anopheles gambiae</i> (Description)						
1	<i>Anopheles gambiae</i> mitochondrion	15,363	L20934	L20934		
2a	<i>Anopheles gambiae</i> str. PEST chromosome 2L (12 parts in a CON entry)	49,364,325	CM000356	CM000356	1438	Proteome
2b	<i>Anopheles gambiae</i> str. PEST chromosome 2R (23 parts in a CON entry)	61,545,105	CM000357	CM000357		
2c	<i>Anopheles gambiae</i> str. PEST chromosome 3L (22 parts in a CON entry)	41,963,435	CM000358	CM000358		
2d	<i>Anopheles gambiae</i> str. PEST chromosome 3R (10 parts in a CON entry)	53,200,684	CM000359	CM000359		
2e	<i>Anopheles gambiae</i> str. PEST chromosome X (13 parts in a CON entry)	24,393,108	CM000360	CM000360		
species <i>Arabidopsis thaliana</i> (Description)						
3a	<i>Arabidopsis thaliana</i> mitochondrion	966,924	Y08501	Y08501	11726	
3b	<i>Arabidopsis thaliana</i> chromosome 1 bottom arm (116 parts in a CON entry)	14,668,883	AE005173	AE005173	13190	Proteome
3c	<i>Arabidopsis thaliana</i> chromosome 1 top arm (148 parts in a CON entry)	14,221,815	AE005172	AE005172		
3d	<i>Arabidopsis thaliana</i> chromosome 3 (331 parts in a CON entry)	23,403,063	BA000014	BA000014		
3e	<i>Arabidopsis thaliana</i> chromosome 4, long arm (78 parts in a CON entry)	14,497,843	AJ270060	AJ270060		
3f	<i>Arabidopsis thaliana</i> chromosome 4, short arm (16 parts in a CON entry)	3,052,119	AJ270058	AJ270058		
3g	<i>Arabidopsis thaliana</i> chromosome 5 (410 parts in a CON entry)	23,810,767	BA000015	BA000015		
3h	<i>Arabidopsis thaliana</i> chloroplast	154,478	AP000423	AP000423	13191	
3i	<i>Arabidopsis thaliana</i> chromosome 2 (265 parts in a CON entry)	19,709,080	At_h2			
species <i>Aspergillus niger</i>						
4a	<i>Aspergillus niger</i> strain N909 mitochondrion	31,103	DQ207726	DQ207726	15772	
4b	<i>Aspergillus niger</i> supercontig SC1, chromosome map 2R (48 parts in a CON entry)	3,625,819	AM270980	AM270980		

II. Genome databases EBI: download genomic data of *A. thaliana*

Integr8 - A.thaliana:

Search for species Search for gene/protein in

Selected species *A.thaliana* [Change scope](#)

Arabidopsis thaliana (cultivar Columbia) - Tax ID: 3702 GAS: [View](#)

Arabidopsis thaliana (mouse ear cress) is a small uninteresting-looking little plant with a rosette of leaves, thin stems and small white flowers, found on the rock exposures of basalt.

Arabidopsis thaliana family, like cabbage these include among weeks from seed to variety of stock col

Integr8 - Download data for A.thaliana:

Search for species Search for gene/protein in

Selected species *A.thaliana* [Change scope](#)

Complete proteome - UniProtKB:

Proteome sets (Fasta/UniProt/XML format)	Gene sets (Fasta/EMBL format)	InterPro hits	GO annotations	Orthologues
Fasta	Fasta	Download	Download	Download
UniProt	EMBL			
XML				

Complete proteome - IPI:

Proteome sets (Fasta/UniProt format)	InterPro hits	Protein cross references	Gene cross references	GOA annotations
Fasta	InterPro	XRets	Gene XRets	Download
UniProt				

Components - UniProtKB:

Genome component	EMBL	Genome Reviews	Proteome sets (Fasta/UniProt format)	Gene sets (Fasta/EMBL format)
Chromosome 2	CT485783	CT485783_GR	Fasta UniProt	Fasta EMBL
Chromosome 4	CT486007	CT486007_GR	Fasta UniProt	Fasta EMBL
Chromosome 3	BA000014	BA000014_GR	Fasta UniProt	Fasta EMBL
Chromosome 5	BA000015	BA000015_GR	Fasta UniProt	Fasta EMBL
Chloroplast	AP000423	AP000423_GR	Fasta UniProt	Fasta EMBL
Chromosome 1	CT485782	CT485782_GR	Fasta UniProt	Fasta EMBL
Mitochondrion	Y08501	Y08501_GR	Fasta UniProt	Fasta EMBL

II. Genome databases

Browsing genomes at NCBI

Genome biology

NCBI provides several genomic biology tools and resources, including organism-specific pages that include links to many web sites and databases relevant to that species. We invite you to explore the links provided on this page.

Map Viewer - genome annotation updates:

Species	Build	Map Viewer Release
Drosophila melanogaster	5.10	November 24, 2009
Drosophila pseudoobscura	2.3	November 24, 2009
Arabidopsis thaliana (mouse-ear cress)	9.1	October 14, 2009
Homo sapiens (human)	37.1	August 4, 2009
Vitis vinifera (wine grape)	IGGP 1	April 7, 2009
Taeniopygia guttata (zebra finch)	1.1	March 5, 2009
Hydra magnipapillata	1.1	January 28, 2009
Physcomitrella patens (moss)	1.1	January 8, 2009
Caenorhabditis elegans (nematode)	WS190	October 10, 2008
Anopheles gambiae (mosquito)	AnamP3.3	October 10, 2008

II. Genome databases

Prokaryotes genomes at NCBI

Complete Microbial Genomes

ENTREZ Genome Project

Search: Genome Project

Organism info Complete genomes Genomes in progress

organism group: -- All --

Tools legend: T - TaxMap; P - ProfTable; C - COG Table; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; F - FTP; R - Publications.

1181 Complete Microbial Genomes selected: [A] - 86, [B] - 1095

RefSeq PID	GPID	Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
49725	30807	Nostoc azollae 0708	B	Cyanobacteria	5.53	38.3			CP002059.1	NC_014248.1	03/06/09	06/16/10	DOE Joint Genome Institute [more]	P L
12997	12997	Acaryochloris marina MBIC11017	B	Cyanobacteria	8.36	47.0	1	9	CP000828.1	NC_009925.1	10/16/07	05/27/10	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine [more]	T P C L S X F R
31129	31129	Acetobacter pasteurianus IFO 3283-01	B	Alphaproteobacteria	3.33	53.1	1	6	AP011121.1	NC_013209.1	08/26/09	04/16/10	Yamaguchi Univ., Japan [more]	T P C L S X F R
31141		Acetobacter pasteurianus IFO 3283-01.42c	B	Alphaproteobacteria	* 3.23	53.1			AP011163		08/26/09		Yamaguchi Univ., Japan [more]	
31131		Acetobacter pasteurianus IFO 3283-03	B	Alphaproteobacteria	* 3.33	53.1			AP011128		08/26/09		Yamaguchi Univ., Japan [more]	
31133		Acetobacter pasteurianus IFO 3283-07	B	Alphaproteobacteria	* 3.33	53.1			AP011135		08/26/09		Yamaguchi Univ., Japan [more]	
32203		Acetobacter pasteurianus IFO 3283-12	B	Alphaproteobacteria	* 3.33	53.1			AP011170		08/26/09		Yamaguchi Univ., Japan [more]	
31135		Acetobacter pasteurianus IFO 3283-22	B	Alphaproteobacteria	* 3.33	53.1			AP011142		08/26/09		Yamaguchi Univ., Japan [more]	
31137		Acetobacter pasteurianus IFO 3283-26	B	Alphaproteobacteria	* 3.33	53.1			AP011149		08/26/09		Yamaguchi Univ., Japan [more]	
31139		Acetobacter pasteurianus	B	Alphaproteobacteria	* 3.33	53.1			AP011156		08/26/09		Yamaguchi Univ.,	

uid=25031 Genome Result

http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&cmd=Retrieve&dopt=Protein+Table&list_uids=25031

Agenda MIC - Calcium ESCO-SA-Silverpeas Genyeastrait INRA-RHselfservice Societe Gén... Logitel Net Apple France Google Maps Wikipedia Divers

uid=25031 Genome Result

Limits Preview/Index History Clipboard Details

Display Protein Table Show 20 Send to

All: 1

Acetobacter pasteurianus *IFO 3283-01, complete genome*

Position: from begin to end Length: from 0 to inf Refresh Reset

Length histogram.
Click on a bar to select length range.

1 2008

2628 protein(s) shown
Legends: Structure

Product Name	Start	End	Strand	Length	Accession	GeneID	Locus	Locus_tag	COG(s)	Protein Clusters	Structure
Holliday junction resolvase RusA	388	813	-	141	YP_003186539.1	8434306	rusA	APA01_00010	COG4570L	CLSK2557702	
phage integrase	1206	2420	-	404	YP_003186540.1	8436902		APA01_00020	COG0582L	CLSK2541448	◆
hypothetical protein APA01_00030	2681	3160	-	159	YP_003186541.1	8434307		APA01_00030	COG1671S	CLSK2402752	
hypothetical protein APA01_00040	3400	3927	+	175	YP_003186542.1	8434308		APA01_00040	-	CLSK933951	
multidrug resistance transporter BcrIC6A	4064	5311	+	415	YP_003186543.1	8434309	bcrIC6A	APA01_00050	COG2814G	CLSK936199	
hypothetical protein APA01_00060	5336	5722	-	128	YP_003186544.1	8436518		APA01_00060	-		
hypothetical protein APA01_00070	5753	5938	-	61	YP_003186545.1	8434310		APA01_00070	COG2835S		◆
Lon-like ATP-dependent protease La	5965	6669	-	234	YP_003186546.1	8434311		APA01_00080	COG2802R	CLSK936150	
thioredoxin	6684	7643	-	319	YP_003186547.1	8434312		APA01_00090	COG3118O	CLSK936149	◆
hypothetical protein APA01_00100	7706	7819	-	37	YP_003186548.1	8434313		APA01_00100	-		

“Omics” databases

Transcriptomics

SMD genome-www5.stanford.edu/cgi-bin/SMD/login.pl
 ArrayExpress www.ebi.ac.uk/microarray/ArrayExpress/arrayexpress.html
 GEO www.ncbi.nlm.nih.gov/geo

Proteomics

SWISS-2DPAGE expasy.org/ch2d
 OPD bioinformatics.icmb.utexas.edu/OPD
 PARIS genome.jouy.inra.fr/paris

Protein-Protein interactions

DIP dip.doe-mbi.ucla.edu
 BIND www.bind.ca
 BRITE www.genome.ad.jp/brite
 STRING string-db.org

Metabolomics

Kegg www.genome.ad.jp/kegg
 Metacyc biocyc.org/meta/
 WIT wit.mcs.anl.gov/WIT2

Comparative genomics resources

UCSC Genome4 Bioinformatics
 Ensembl
 MapViewer
 VISTA Genome Browser
 Comparative Regulatory Genomics
 GALA
 EnsMart
 PipMaker and MultiPipMaker
 VISTA server
 MAVID server
 zPicture server
 rVISTA server
 COGs
 MOSAIC

genome.ucsc.edu
www.ensembl.org
www.ncbi.nlm.nih.gov/mapview
pipeline.lbl.gov
corg.molgen.mpg.de
www.bx.psu.edu
www.ensembl.org/EnsMart
www.bx.psu.edu
www-gsd.lbl.gov/vista
baboon.math.berkeley.edu/mavid
zpicture.dcode.org
rvista.dcode.org
www.ncbi.nlm.nih.gov/COG
genome.jouy.inrafr/mosaic

Databases of Motifs and Mobile elements

Regulation motifs

Transfac www.biobase-international.com/pages/index.php?id=transfac
 RegulonDB regulondb.ccg.unam.mx

Protein motifs

Interpro www.ebi.ac.uk/interpro
 Pfam www.sanger.ac.uk/Software/Pfam

Mobile elements

Isfinder www-is.biotoul.fr/is.html
 ACLAME aclame.ulb.ac.be

Repeat elements

Repbase www.girinst.org/repbase/index.html
 CRISPRdb crispr.u-psud.fr/crispr

Genome databases: Ensembl, UCSC, MapViewer

What are genome databases?

- **Genome databases contain, well, genomic information collected from many sources.**
 - Genome assembly
 - Gene predictions
 - Known genes, mRNA, ESTs, proteins
 - Genetic maps, markers and polymorphisms
 - Gene expression and phenotypes
 - Annotations
 - Interspecies homologues

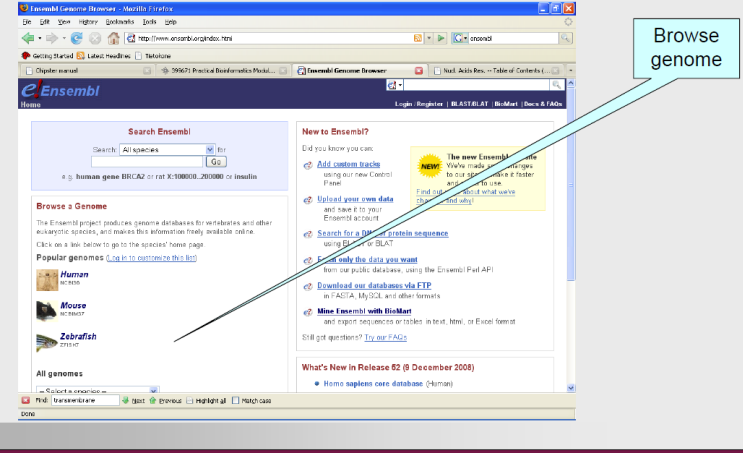
Why genome databases?

- **Genome structure**
- **Gene identification**
- **Complete catalog or blueprint**
- **Rapid identification of proteins**
- **Genetic, transcriptome, proteome analysis**
- **Comparative genomics**

Some considerations

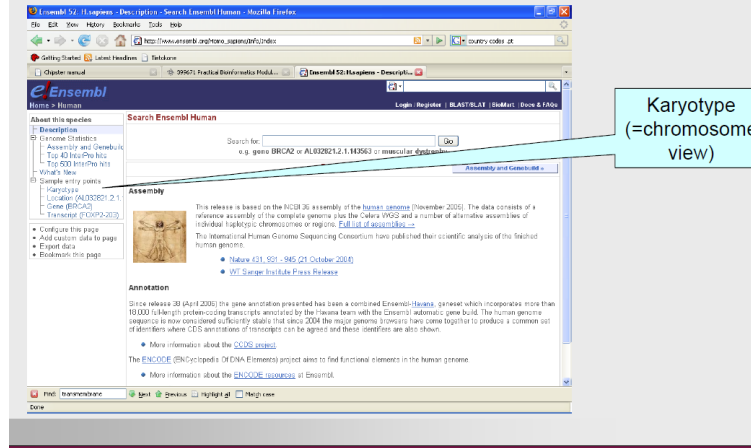
- **Selection of the database**
 - Organism content
 - Speed (MapViewer can be slow)
- **Organism specific databases can be more up-to-date than general databases**
- **Genome databases are not a one stop shop for all information, other databases like EMBL and UniProt are still needed**

Ensembl front page



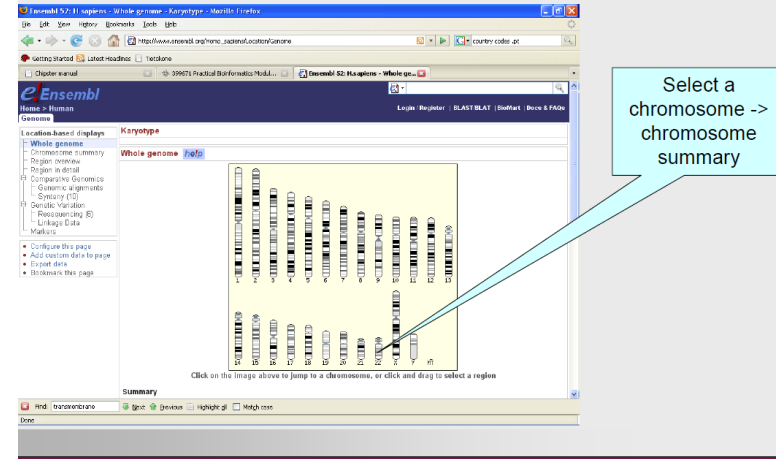
Browse genome

Explore the genome



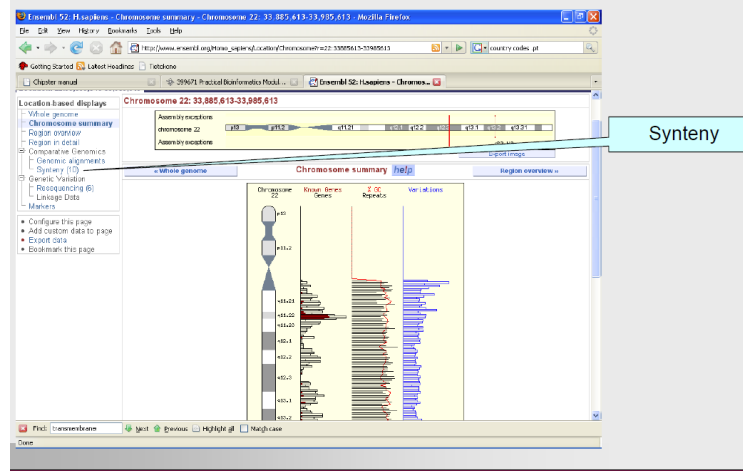
Karyotype (=chromosome view)

Explore chromosomes



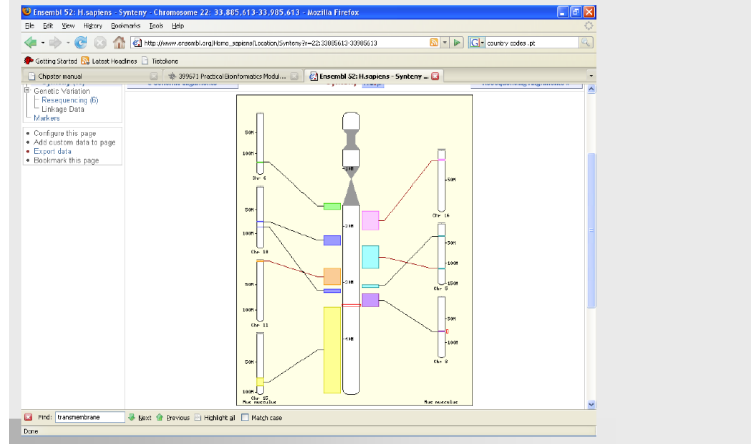
Select a chromosome -> chromosome summary

Chromosome summary



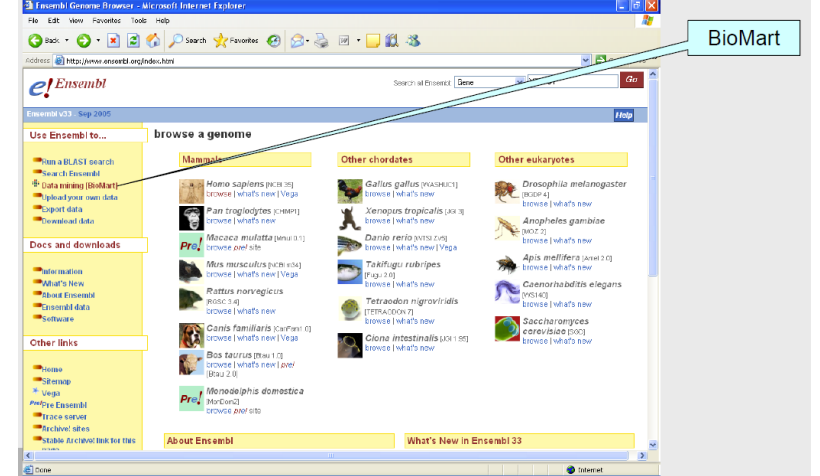
Synteny

Synteny View



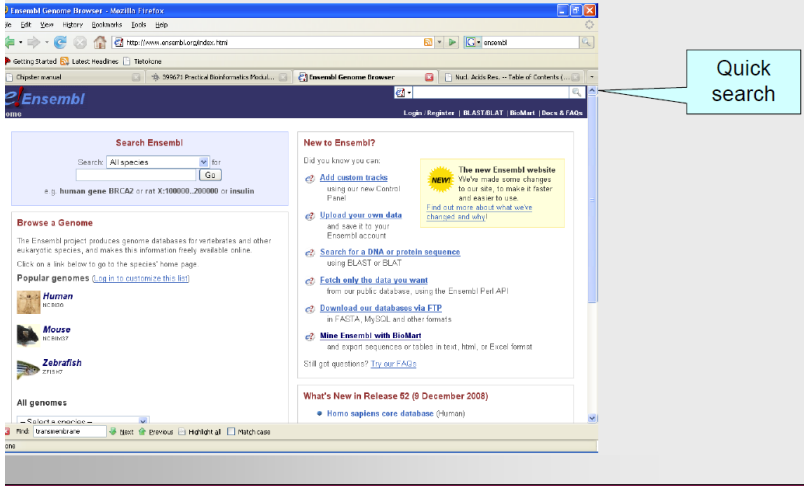
Synteny

Ensembl front page



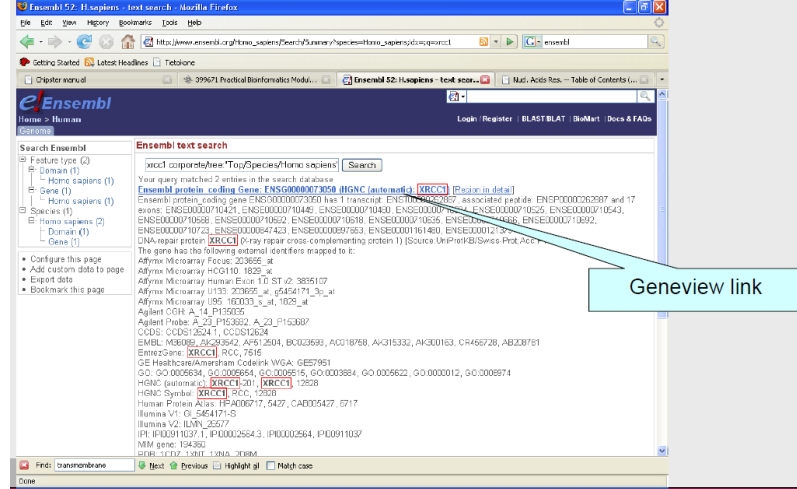
BioMart

Ensembl front page



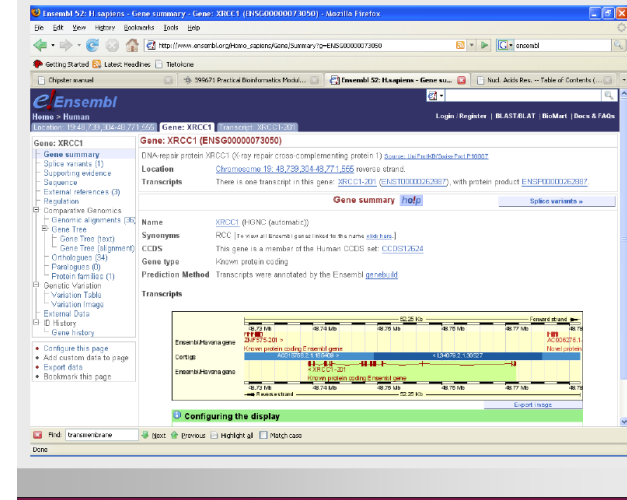
Quick search

Quick search results

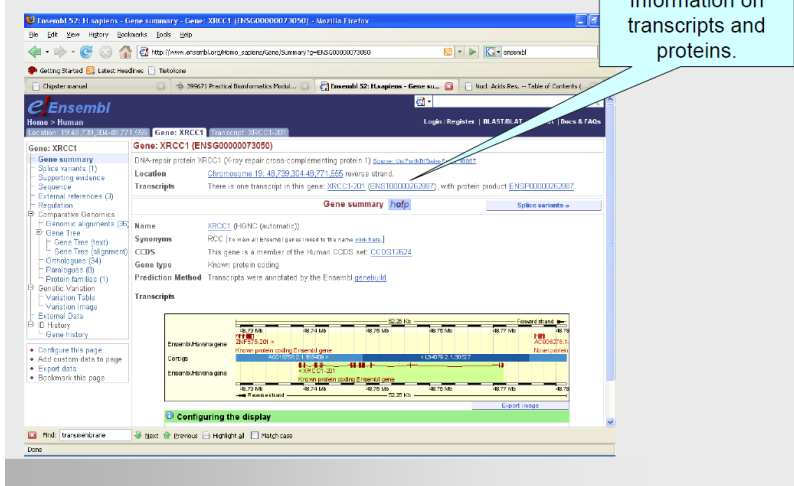


Geneview link

Gene View

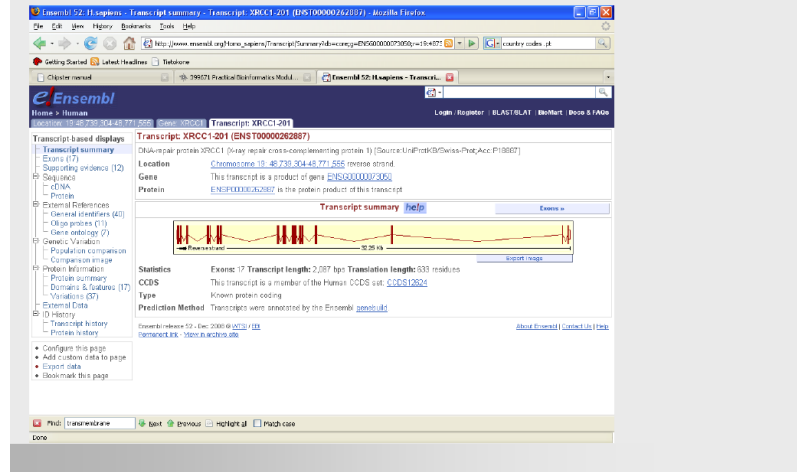


Gene View

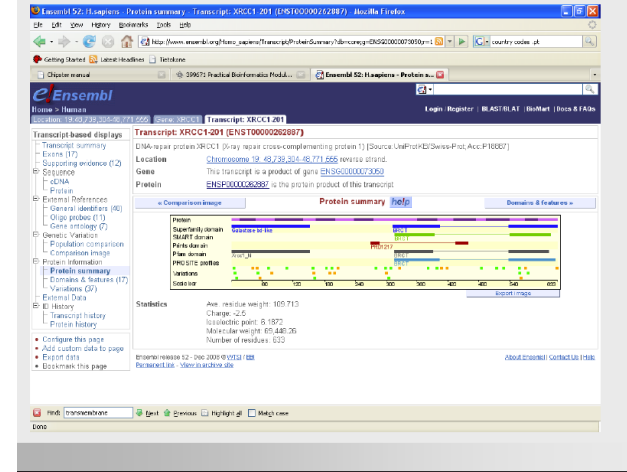


Information on transcripts and proteins.

Transcript info



Protein summary



Variation (SNP) view – from Gene tab

This screenshot shows the Ensembl Variation view for the XRC1 gene. The main panel displays a 'Variation Table' with columns for Variation ID, Position, and other details. Below the table is a 'Variation Image' showing a genomic track with various annotations and a 'Variation Image' link. The left sidebar contains navigation options like 'Gene summary', 'Splice variants', and 'Variation Table'. The top navigation bar includes 'Home', 'Human', and 'Login'.

SNPView (link from protein summary)

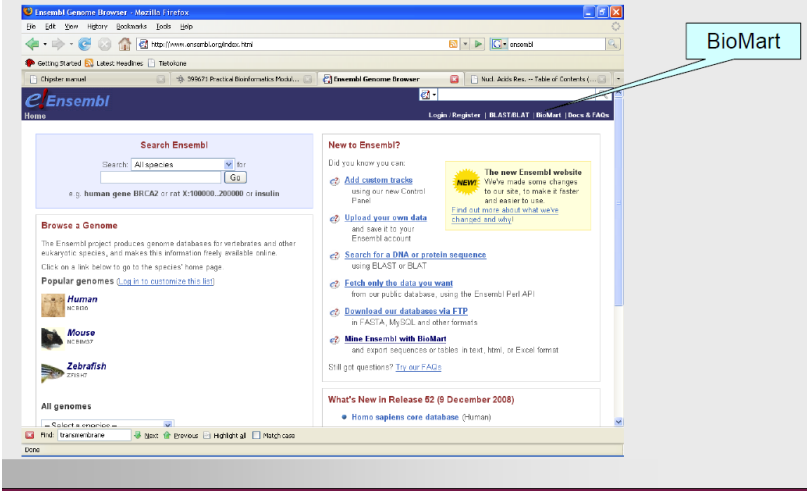
This screenshot shows the Ensembl SNPView for rs25489. The main panel displays 'Variation summary' for rs25489, including 'Variation type' (SNP), 'Synonyms', and 'Location'. Below this is a 'Variation image' showing a genomic track with various annotations and a 'Variation image' link. The left sidebar contains navigation options like 'Variation summary', 'Validation status', and 'Flanking Sequence'. The top navigation bar includes 'Home', 'Human', and 'Login'. A callout box points to the 'Linkage disequilibrium data' section, which shows 'Linkage disequilibrium data per population' and 'Linkage disequilibrium data per population'.

View linkage disequilibrium in the population (LDView).

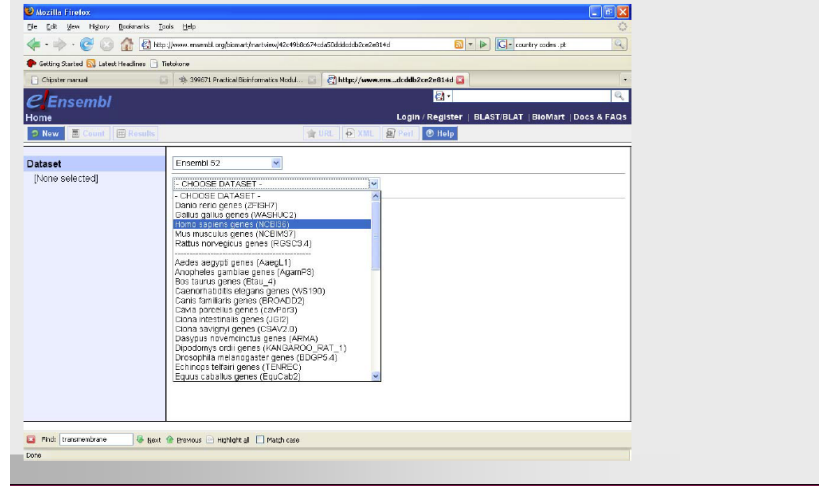
LDView

This screenshot shows the Ensembl LDView for the XRC1 gene. The main panel displays a 'Linkage Disequilibrium Data' plot for XRC1. The plot shows 'Variants' on the x-axis and 'Genotype variations' on the y-axis. The plot is color-coded by 'Variation legend' (Transcript, Non-synonymous coding, Synonymous coding, Splice site SNP) and 'Tagged SNPs'. Below the plot is a 'LD (r²)' heatmap showing the correlation between variants. The top navigation bar includes 'Home', 'Human', and 'Login'. The bottom navigation bar includes 'Find', 'Transcript', 'Gene', 'Previous', 'Highlight all', and 'Match case'.

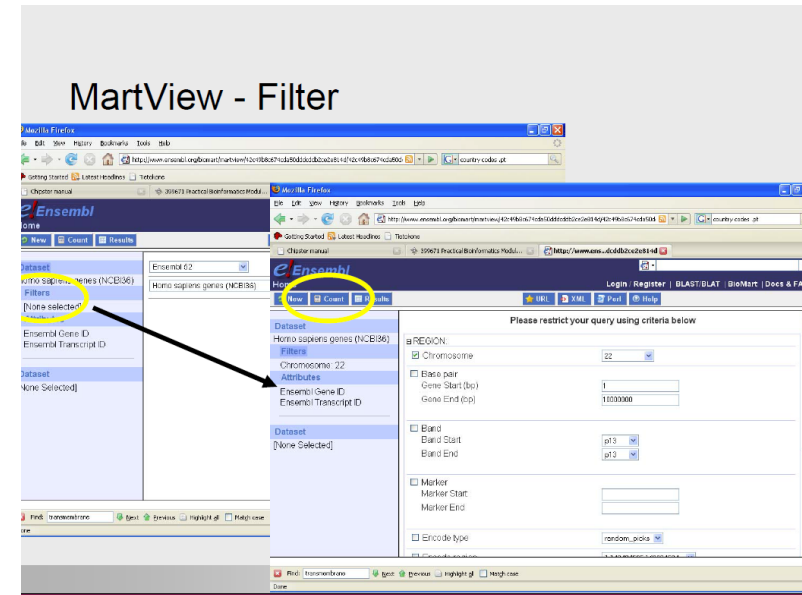
Ensembl front page



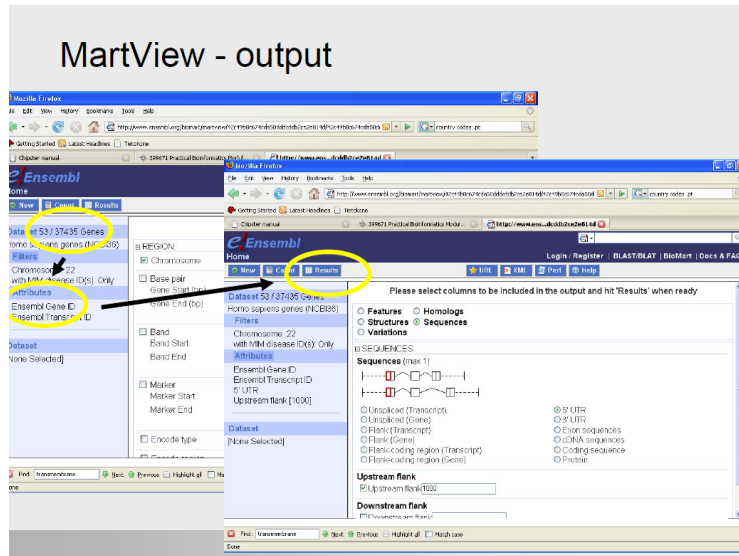
MartView – select genome



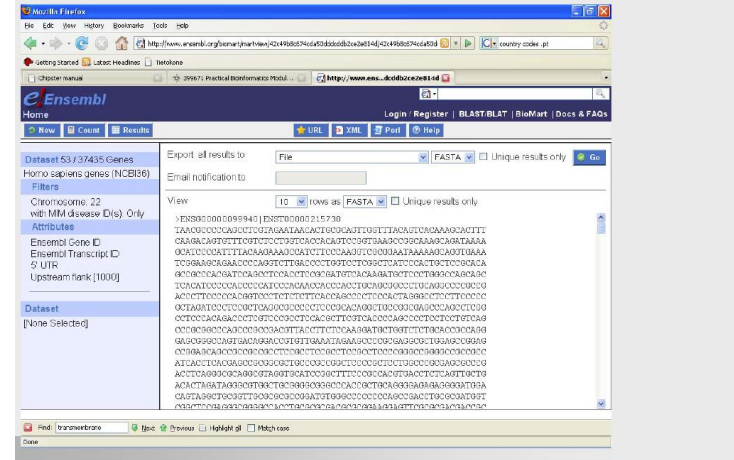
MartView - Filter



MartView - output



MartView



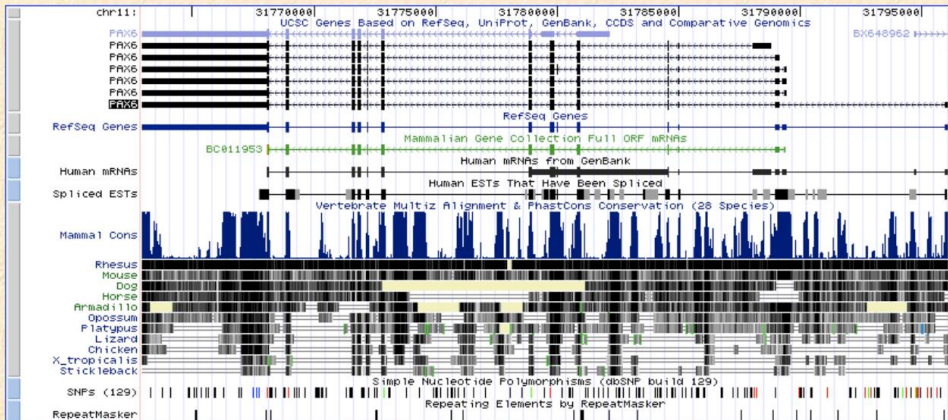
Human gene Pax6 aligned with Vertebrate genomes

UCSC Genome Browser on Human Mar. 2006 Assembly

move <<<< << < > >> >>>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr11:31,762,916-31,796,085 jump clear size 33,170 bp. configure

chr11 (p13) p15.4 p13 p12 q14.1 q21 q22.3 25



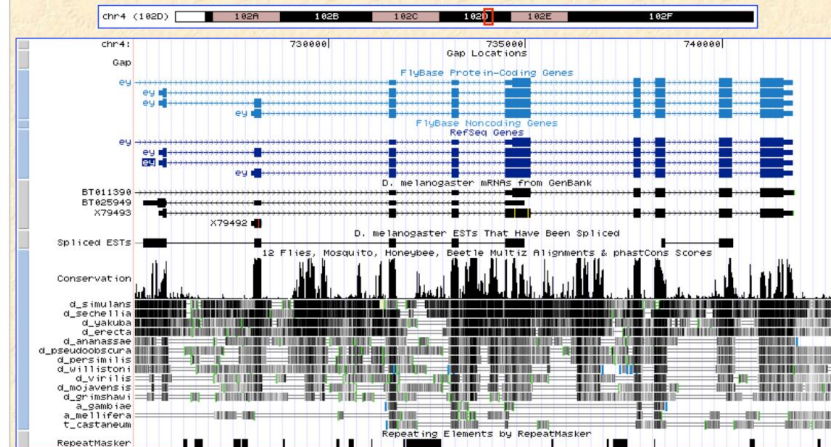
move start Click on a feature for details. Click on base position to zoom in around cursor. Click move end
 < 2.0 > gray/blue bars on left for track options and descriptions. < 2.0 >

Drosophila gene eyeless (homolog to Pax6) aligned with Insect genomes

UCSC Genome Browser on D. melanogaster Apr. 2006 Assembly

move <<<< << < > >> >>>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr4:725,098-742,933 jump clear size 17,836 bp. configure



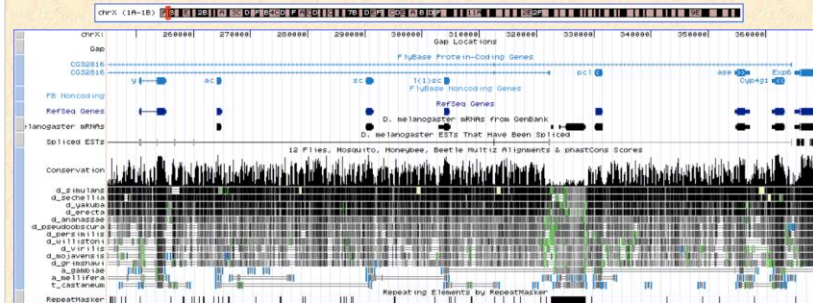
move start Click on a feature for details. Click on base position to zoom in around cursor. Click move end
 < 2.0 > gray/blue bars on left for track options and descriptions. < 2.0 >

Drosophila 120kb chromosomal region covering the Achaete-Scute Complex

UCSC Genome Browser on D. melanogaster Apr. 2006 Assembly

move <<<< << < > >> >>>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chrX:245,000-370,000 jump clear size 125,001 bp. configure



move start Click on a feature for details. Click on base position to zoom in around cursor. Click move end
 < 2.0 > options and descriptions. < 100 >

Comparative genomics

Integr8 - access to complete genomes and proteomes

<http://www.ebi.ac.uk/integr8/>



Integr8 - genome summaries

<http://www.ebi.ac.uk/integr8/>



- Home
- local help
- Integr8 News
- Focal Point archive
- Latest Species
- Browse Species
- Inquisitor status
- BioMart
- Proteomes and Genomes FASTA
- About Integr8
- Publications
- Integr8 Web service
- Genome Reviews
- IPI

EBI > Databases > Integr8

Integr8 : Access to complete genomes and proteomes

Search for species Search for gene/protein in

scope **Bacteria, Archaea, Eukaryota** [Change scope](#)

The **Integr8** web portal provides easy access to integrated information about deciphered genomes and their corresponding proteomes. Available data includes DNA sequences (from databases including the EMBL Nucleotide Sequence Database, Genome Reviews, and Ensembl); protein sequences (from databases including the UniProt Knowledgebase and IPI); statistical genome and proteome analysis (performed using InterPro, CluStr, and GOA); and information about orthology, paralogy, and synteny.

Integr8 data can also be accessed via the [Integr8 FTP](#) site.

New to Integr8? The [user guide](#) will show you how to make the most of the data provided by Integr8. Alternatively, you may choose to [start browsing the data](#). We value your feedback! Please [send us your comments](#).

News

A complete list of Integr8 species and their proteome status can be found on the [current status](#) page of the Integr8 documentation.

This release of Integr8 (release 89) was built from UniProt release 14.5 and InterPro release 18.0 and was released on Tue, Nov 25, 2008.

The summary chart below shows the types of species currently held within Integr8.

Click on the chart to browse species in Integr8 by taxonomic classification.

History

Rank	Species	Count
(1)	Homo sapiens	27048
(2)	Mus musculus	20426
(3)	Saccharomyces cerevisiae ATCC 204508	19672
(4)	Escherichia coli DSM 5911	11607
(5)	Escherichia coli K12	11329
(6)	Schizosaccharomyces pombe	9532
(7)	Drosophila melanogaster	7696
(8)	Rattus norvegicus	5886
(9)	Caenorhabditis elegans	4352
(10)	Arabidopsis thaliana	4205

- Home
- local help
- Integr8 News
- Focal Point archive
- Latest Species
- Browse Species
- Inquisitor status
- BioMart
- Proteomes and Genomes FASTA
- About Integr8
- Publications
- Integr8 Web service
- Genome Reviews
- IPI

EBI > Databases > Integr8

Integr8 : O.sativa Nipponbare Genome Statistics:

Search for species Search for gene/protein in

Selected species **O.sativa Nipponbare** [Change scope](#)

Component name	Protein count	Type	Length (bp)	Av. CDS Length	GC content	CDS coverage	Gene count
Chromosome 1	3665	---	43201740	1134.655	43.0%	11%	3646
Chromosome 2	3070	---	35954743	1200.630	43.3%	10%	3004
Chromosome 3	3362	---	36192742	1130.954	43.7%	11%	3345
Chromosome 4	2566	---	35498469	1211.117	44.2%	8%	2385
Chromosome 5	2150	---	2973217	1158.593	44%	6%	2163
Chromosome 6	2145	---	3073186	1238.862	43.6%	8%	2143
Chromosome 7	2690	---	29644043	1192.297	43.5%	8%	2082
Chromosome 8	1801	---	28434780	1133.421	43.4%	8%	1707
Chromosome 9	1464	---	22390651	1177.715	43.5%	6%	1409
Chromosome 10	1436	---	22959606	1205.62	43.6%	6%	1434
Chromosome 11	1581	---	28388948	1281.537	42.9%	7%	1575
Chromosome 12	1601	---	27599993	1201.039	43%	6%	1405
Mitochondrion	53	---	490520	624.266	43.9%	9%	53
Chloroplast	88	---	134551	708.524	35%	33%	88
Total	20330						85

Protein number per component:

(Hover mouse over sections of chart to display protein number)

Amino acid composition:

Protein length distribution:

Triplet usage:

Integr8 - clusters of orthologous genes (COGs)

<http://www.ebi.ac.uk/integr8/>

- Home
- local help ⓘ
- Integr8 News
- Focal Point archive
- Latest Species
- Browse Species
- H.sapiens
 - Literature
 - Genome Statistics
 - Proteome Analysis
 - Downloads
 - Taxonomy
- Gene search results
- Integr8or
- Inquisitor status
- BioMart
- Proteomes and Genomes FASTA
- About Integr8

EBI > Databases > Integr8

Integr8 : Integr8or

Search for species Go! Search for gene/protein in H.sapiens Go!

Selected species **H.sapiens** gene **PAX6** [Change scope](#)

Gene
Results
Context
History

Taxonomic spread for Putative ORthologous Cluster : 99724 Name: Paired box protein Pax-6 Show/Hide Tree ▲

Loading tree...

Similar sequences in other species ⓘ 9 results

Members of the displayed cluster are represented in same color (non white)

Select genes to display Synteny Align

Protein	Chromosome	Organism	PORC ID	Select
Paired box protein Pax-6	Chromosome 2	M.musculus	99724	*
Paired box protein Pax-6	Chromosome 15	B.taurus	99724	*
Paired box protein Pax-6	Chromosome 3	R.norvegicus	99724	*
Paired box protein Pax[Zf-a]	Chromosome 25	D.erio	99724	*
Chromosome 5 SCAF14773, whole genome shotgun sequence	Unassembled WGS sequence	T.nigroviridis	99724	N/A
Paired box protein Pax-6	Chromosome 5	G.gallus	99724	*
CG11186	Chromosome 4	D.melanogaster	99724	N/A
MAB-18	Chromosome X	C.elegans	99724	*
Paired box protein pax-6	Unassembled WGS sequence	A.aegypti	99724	N/A

Integr8 - clusters of paralogous genes

<http://www.ebi.ac.uk/integr8/>

- Home
- local help ⓘ
- Integr8 News
- Focal Point archive
- Latest Species
- Browse Species
- H.sapiens
 - Literature
 - Genome Statistics
 - Proteome Analysis
 - Downloads
 - Taxonomy
- Gene search results
- Integr8or
- Inquisitor status

EBI > Databases > Integr8

Integr8 : Integr8or

Search for species Go! Search for gene/protein in H.sapiens Go!

Selected species **H.sapiens** gene **PAX6** [Change scope](#)

Gene
Results
Context
History

Similar sequences in H.sapiens ⓘ 8 results

Select genes to display Synteny Align

Protein	Chromosome	Organism	Select
Paired box protein Pax-7	Chromosome 1	H.sapiens	*
Paired box protein Pax-3	Chromosome 2	H.sapiens	*
Paired box protein Pax-4	Chromosome 7	H.sapiens	*
Paired box protein Pax-2	Chromosome 10	H.sapiens	*
Paired box protein Pax-5	Chromosome 9	H.sapiens	*
Paired box protein Pax-8	Chromosome 2	H.sapiens	*
Paired box protein Pax-1	Chromosome 20	H.sapiens	*
Paired box protein Pax-9	Chromosome 14	H.sapiens	*

Genome resources challenges

Some biological challenges :

- **Genome annotation:** Merging automated, experimental and curated information : “reference genomes” vs “draft genomes”
- **Dealing with multiple genomes concerning:** individuals, strains and related species
- **Linking polymorphisms with phenotypes and functional studies**

Some bioinformatics challenges :

- **Data archiving :** format, volume and standardization problems
- **Data integration:** physical, virtual, semantical
- **Data visualization** of large volume of data in a visually intuitive format

Conclusion

- **NGS (Next Generation Sequencing) technologies** provide more and more data at ever lower cost.
- **Diversification of projects:** de novo sequencing, re-sequencing, metagenomics, RNAseq,...

Example: the 1000 genomes project www.1000genomes.org

- We will have to deal with increasing amounts of sequencing data and **it is still a challenging task to provide adequate structures to produce, store and analyze data**

Table 1. A variety of resources for the study of collections of prokaryotic genomes

Excellent sources of information about a wider range of databases and web services are the special database and web server issues of *Nucleic Acids Research*, which are published every January and July, respectively (<http://nar.oupjournals.org/>).

PROKARYOTIC GENOMIC RESOURCES	
Monitoring completed and ongoing genome projects	
Genomes Online Database (GOLD) http://www.genomesonline.org	Provides access to lists of complete and ongoing genome projects from prokaryotes and eukaryotes
Primary international databases of complete genome sequences	
DNA Database of Japan (DDBJ) http://gib.genes.nig.ac.jp/	Genomes at DDBJ in the Genome Information Broker system
European Bioinformatics Institute (EBI) http://www.ebi.ac.uk/genomes/	Genomes at EBI
National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome	Genomes at NCBI in the Entrez Genomes system
Specialized databases	
A Systematic Annotation Package for Community Analysis of Genomes (ASAP) https://asap.ahabs.wisc.edu/annotation/php/logon.php	Genome sequences, annotations and experimental data for multiple organisms plus an interface for direct community contributions
Molligen http://cbi.labri.fr/outils/molligen/	Website dedicated to mollicute genomes allowing BLAST searching, whole-genome alignment
Oral Pathogens database http://www.oralgen.lanl.gov/	Databases of oral pathogens, bacterial and viral
Pathema http://www.tigr.org/pathema/index.shtml	In-depth curatorial analysis of pathogen genomes
STDGen and the Oral Pathogens database http://www.stdgen.lanl.gov/	Databases of genomes responsible for sexually transmitted diseases
Comparative genomic databases	
KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/	Enzyme and pathway information about complete genomes
Comprehensive Microbial Resource (CMR) http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl	Provides access to a wide range of information and analyses about all complete bacterial genomes
Integrated Microbial Genomes (IMG) http://img.jgi.doe.gov/v1.0/main.cgi	Facilitates the visualization and exploration of genomes from a functional and evolutionary perspective
Microbial Genome Database for Comparative Analysis (MBGD) http://mbgd.genome.ad.jp/	Provides orthologue identification, paralogue clustering, motif analysis and gene order data
Virulogenome http://www.vge.ac.uk/index.html	Access to complete and incomplete genomes, including Artemis applet and ACT comparisons
Genomic feature databases	
Clusters of Orthologous Genes (COGs) http://www.ncbi.nlm.nih.gov/COG/	Individual proteins or groups of paralogues from at least three lineages corresponding to ancient conserved domains
FusionDB http://igs-server.cnrs-mrs.fr/FusionDB/	A database of bacterial and archaeal gene fusion events
Genome Atlas http://www.cbs.dtu.dk/services/GenomeAtlas/	Visualization of features within large regions of DNA; users can upload GenBank files to create custom plots
High-quality Automated and Manual Annotation of microbial Proteomes (HAMAP) http://www.expasy.org/sprot/hamap/	HAMAP families are a collection of orthologous microbial protein families, generated manually by expert curators
Genome Reviews http://www.ebi.ac.uk/GenomeReviews/	Up-to-date, standardized and comprehensively annotated view of the genomes
Homologous Sequences in Complete Genomes Database http://pbil.univ-lyon1.fr/databases/hogenom.html	Database of homologous genes and access to phylogenetic trees
Merops http://merops.sanger.ac.uk/	Information resource for peptidases and the proteins that inhibit them

Table 1. cont.

PROKARYOTIC GENOMIC RESOURCES	
ORFanage http://www.cs.bgu.ac.il/~nomsiew/ORFans/	Access to singleton, paralogous and orthologous ORFans in bacterial genomes
OrphanMINE http://www.genomics.ceh.ac.uk/orphan_mine/	Database of bacterial proteomes with access to lists of orphans that can be filtered by a variety of criteria
Pathogenomics http://www.pathogenomics.bc.ca/IslandPathExamples.html	Identification of horizontally transferred genes and genomics islands, including pathogenicity islands
SEED http://theseed.uchicago.edu/FIG/index.cgi	Expert curation of genomic subsystems, or sets of functionally or phenotypically related genes
TransportDB http://www.membranetransport.org/	Database describing the predicted cytoplasmic membrane transport proteins
tRNADB http://lowelab.ucsc.edu/GtRNADB/	Genomic tRNA database which contains tRNA identifications made by the program tRNAscan-SE
Pathway and protein interaction databases	
BioCyc http://www.biocyc.org/	A collection of curated databases each of which describes the genome and metabolic pathways of a single organism
MetaCyc http://metacyc.org/	A database of nonredundant, experimentally elucidated metabolic pathways
STRING http://string.embl.de/	A database of known and predicted protein-protein interactions
Multiple genome alignment tools	
A Genome Comparison Tool (ACT) http://www.sanger.ac.uk/Software/ACT/	A DNA sequence comparison viewer (usually BLASTN or tBLASTX) based on the Artemis genome visualization tool
Mauve http://gel.ahabs.wisc.edu/mauve/	Multiple genome alignments in the presence of large-scale evolutionary events
Multi-LAGAN http://lagan.stanford.edu/lagan_web/index.shtml	One of several packages in the LAGAN tool set for multiple alignment of genomes
MultiPipMaker http://pipmaker.bx.psu.edu/pipmaker/	Summarizes similarity between multiple sequences using 'percent identity plots' (Pips)
Multiple Genome Aligner (MGA) http://bibiserv.techfak.uni-bielefeld.de/mga/	Computation of multiple genome alignments of large, closely related DNA sequences
Phylogenomics	
PyPhy http://www.cbs.dtu.dk/staff/thomas/pyphy/	Automatic, large-scale reconstructions of phylogenetic relationships of complete microbial genomes
Phylogenomic Display of bacterial genes (Phydbac) http://igs-server.cnrs-mrs.fr/phydbac/	Web interactive tool that displays phylogenomic profiles of bacterial protein sequences
Visualization of multiple genomes	
EnteriX http://globin.cse.psu.edu/enterix/	Visualization tools for bacterial genome alignments
Multiple Genome Navigator (MuGeN) http://www-mig.jouy.inra.fr/bdsi/MuGeN/	Tool for visual exploration of features of multiple genomes
Genomic metadata	
CMR's Genome Properties http://www.tigr.org/Genome_Properties/	Numerous attributes whose status can be described by numerical values or controlled vocabulary terms
GenomeMine http://www.genomics.ceh.ac.uk/GMINE/	Database of information about all complete genomes
Integr8 www.ebi.ac.uk/integr8/	Access to species descriptions, literature, statistical analysis and summary information about proteomes
NCBI Genome Projects http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj	Organism-specific overviews that function as portals for all projects in the database
Systematic Analysis of Completely Sequenced Organisms (SASCO) http://www.pasteur.fr/~tekaia/sasco.html	Information on base composition, amino acid composition, ancestral duplication, ancestral conservation and organisms' classification

Proteomske podatkovne baze

SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny

Derek Wilson^{1,*}, Ralph Pethica², Yiduo Zhou², Charles Talbot², Christine Vogel³, Martin Madera², Cyrus Chothia¹ and Julian Gough²

SUPERFAMILY provides structural, functional and evolutionary information for proteins from all completely sequenced genomes, and large sequence collections such as UniProt. Protein domain assignments for over 900 genomes are included in the database, which can be accessed at <http://supfam.org/>. Hidden Markov models based on Structural Classification of Proteins (SCOP) domain definitions at the superfamily level are used to provide structural annotation. We recently produced a new model library based on SCOP 1.73. Family level assignments are also available. From the web site users can submit sequences for SCOP domain classification; search for keywords such as superfamilies, families, organism names, models and sequence identifiers; find over- and underrepresented families or superfamilies within a genome relative to other genomes or groups of genomes; compare domain architectures across selections of genomes and finally build multiple sequence alignments between Protein Data Bank (PDB), genomic and custom sequences. Recent extensions to the database include InterPro abstracts and Gene Ontology terms for superfamilies, taxonomic visualization of the distribution of families across the tree of life, searches for functionally similar domain architectures and phylogenetic trees. The database, models and associated scripts are available for download from the ftp site.

SUPERFAMILY is a database of structural and functional annotation for all proteins and genomes.

The SUPERFAMILY annotation is based on a collection of hidden Markov models, which represent structural protein domains at the SCOP superfamily level. A superfamily groups together domains which have an evolutionary relationship. The annotation is produced by scanning protein sequences from **over 2,478 completely sequenced genomes** against the hidden Markov models.

SUPERFAMILY 1.75 including a domain-centric gene ontology method

David A. de Lima Morais^{1,*}, Hai Fang¹, Owen J. L. Rackham¹, Derek Wilson², Ralph Pethica¹, Cyrus Chothia² and Julian Gough^{1,*}

The SUPERFAMILY resource provides protein domain assignments at the structural classification of protein (SCOP) superfamily level for over 1400 completely sequenced genomes, over 120 metagenomes and other gene collections such as UniProt. All models and assignments are available to browse and download at <http://supfam.org>. A new hidden Markov model library based on SCOP 1.75 has been created and a previously ignored class of SCOP, coiled coils, is now included. Our scoring component now uses HMMER3, which is in orders of magnitude faster and produces superior results. A cloud-based pipeline was implemented and is publicly available at Amazon web services elastic computer cloud. The SUPERFAMILY reference tree of life has been improved allowing the user to highlight a chosen superfamily, family or domain architecture on the tree of life. The most significant advance in SUPERFAMILY is that now it contains a domain-based gene ontology (GO) at the superfamily and family levels. A new methodology was developed to ensure a high quality GO annotation. The new methodology is general purpose and has been used to produce domain-based phenotypic ontologies in addition to GO.

Sequence analysis

Advance Access publication January 21, 2014

InterProScan 5: genome-scale protein function classification

Philip Jones^{1,2}, David Binns¹, Hsin-Yu Chang¹, Matthew Fraser¹, Weizhong Li¹, Craig McAnulla¹, Hamish McWilliam¹, John Maslen^{1,2}, Alex Mitchell^{1,*}, Gift Nuka¹, Sebastien Pesseat¹, Antony F. Quinn¹, Amaia Sangrador-Vegas¹, Maxim Scheremetjew¹, Siew-Yit Yong¹, Rodrigo Lopez¹ and Sarah Hunter^{1,*}

ABSTRACT

Motivation: Robust large-scale sequence analysis is a major challenge in modern genomic science, where biologists are frequently trying to characterize many millions of sequences. Here, we describe a new Java-based architecture for the widely used protein function prediction software package InterProScan. Developments include improvements and additions to the outputs of the software and the complete reimplementations of the software framework, resulting in a flexible and stable system that is able to use both multiprocessor machines and/or conventional clusters to achieve scalable distributed data analysis. InterProScan is freely available for download from the EMBI-EBI FTP site and the open source code is hosted at Google Code.

Availability and implementation: InterProScan is distributed via FTP at <ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/> and the source code is available from <http://code.google.com/p/interproscan/>.

Reorganizing the protein space at the Universal Protein Resource (UniProt)

The UniProt Consortium^{1,2,3,4,*}

The mission of UniProt is to support biological research by providing a freely accessible, stable, comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and querying interfaces. UniProt is comprised of four major components, each optimized for different uses: the UniProt Archive, the UniProt Knowledgebase, the UniProt Reference Clusters and the UniProt Metagenomic and Environmental Sequence Database. A key development at UniProt is the provision of complete, reference and representative proteomes. UniProt is updated and distributed every 4 weeks and can be accessed online for searches or download at <http://www.uniprot.org>.

Representative proteomes. There are hundreds of complete proteomes not included in the UniProt Reference Proteomes and this number is expected to increase many fold with sequences from new organisms as well as additional isolates and strains of existing organisms. This flood of new proteomes will decrease the sensitivity of sequence and text searches. To help cope with this, we are working on a computationally derived set of RPs. A RP is the proteome that can best represent all the proteomes in its group in terms of the majority of the sequence space and annotation (7). Each RP is selected

Bibliografske/Literaturne podatkovne baze

Literature Databases

- PubMed / MEDLINE

- Database of citations and abstracts for biomedical literature



- OMIM (Online Mendelian Inheritance in Man) [Glucokinase]

- Catalog of human genes and genetic disorders with textual information and copious links to scientific literature



- Google Scholar



- CiteXplore

- combines literature search with text mining tools for biology.



- Arxiv

- Open access to 601,910 e-prints in Physics, Mathematics, Computer Science, Quantitative Biology, Quantitative Finance and Statistics

Bibliographic databases

- Pubmed: www.pubmed.org

- Comprises more than 19 million citations for biomedical literature from MEDLINE, life science journals, and online books.
- Citations may include links to full-text content from PubMed Central and publisher web sites.

- ISI Web of Science: wokinfo.com/products_tools/multidisciplinary/webofscience

- A research platform providing access to the world's leading citation databases
- Web of Science information is carefully evaluated and selected.

- Free access journals: authors pay to get the papers free

- The Biomed Central BMC initiative: www.biomedcentral.com
- The Public Library Of Sciences initiative: www.plos.org

Taxonomy

- UniProt taxonomy [homo sapiens]

- Organisms are classified in a hierarchical tree structure.
- next to manually verified organism names, external links, organism strains and viral host information is provided.

- NCBI taxonomy [homo sapiens]

Mnemonic	HUMAN
Taxon identifier	9606
Scientific name	Homo sapiens
Common name	Human
Synonym	-
Other names	> man
Rank	Species
Lineage	> cellular organisms > Eukaryota > Fungi/Metazoa group > Metazoa > Eumetazoa > Bilateria > Coelomata > Deuterostomia > Chordata > Craniata > Vertebrata



Problemi oz. težave bioloških podatkovnih baz

Issues for biological databases

- Dealing with biological complexity
- Data content
 - Coverage
 - Information content
- Data quality
 - Data structure
 - Consistency
- Query capabilities
- Interfaces
 - User interfaces
 - Programmatic interfaces
- Annotation
- Funding

Towards biological complexity

- The main databases currently available are focussed on one type of molecular entity : nucleic sequences, proteins, compounds, ...
- This type of organization is very convenient as far as the information to be represented is simple (e.g. DNA sequences, structures of small molecules and macromolecules).
- It becomes more difficult if we want to represent
 - the interactions between biological objects,
 - the integration of various elements in a biological process (metabolic pathways, protein interaction networks, regulatory networks, ...)
 - complex concepts such as "biological function"

Data content

- Scope of the database
 - types of biological objects represented
- Number of entries
 - coverage of the current knowledge
- Information content
 - Level of detail in the description of the biological objects
- References to the source of information

Data quality

- Data Consistency
 - always use the same name to indicate the same object
 - (this seems trivial, but its is unfortunately still not always the case)
 - event better: define an ID for each objects, and allow to retrieve it by any of its synonyms
 - spelling mistakes
- Data Structuration
 - distinct fields for distinct attributes of the biological objects
- Reliability
 - Evidences ? Level of confidence ?
 - Assingnation of function by similarity
 - recursive process → propagation of errors

Query capabilities

- Browsing (click and read)
- Simple search
 - select records with some constraints
- More elaborate search
 - select specific fields of some records with constraints on some fields (~SQL SELECT)
- Complex querying
 - ability to return an answer that results from a "live" computation, and was not part of any record of the dabatase

Interfaces

- User interfaces
 - user-friendly
 - convenient browsing
 - intuitive query forms
 - visualization (graphical output)
- Programmatic interfaces
 - communication with external programs:
 - other databases (concept of distributed database)
 - analysis tools

Annotation

- Problem
 - The flow of available data is increasing exponentially
- Strategies
 - internal curators
 - selected external experts
 - public submission
 - computer-based extraction of information from biological texts

Funding

- **Public funding**
 - Problem: easier to obtain public funds for creating a new database than for maintaining or expanding existing resources
- **Private funding**
 - Industrial companies are
 - ready to invest in good data and good query capabilities
 - interested by academic expertise
- **Solutions**
 - All users pay (per query for example)
 - Note: academic users are anyway funded by public funds
 - Hybrid solution
 - access is free for academic users, not for companies
 - companies can buy the whole database and install it in-house (+ add their own private data)
 - academia-industry interface is often ensured by a spinoff company

A final rant ...

Bio-databases: A short word on problems

- Even today we face some key limitations
 - There is no standard format
 - Every database or program has its own format
 - There is no standard nomenclature
 - Every database has its own names
 - Data is not fully optimized
 - Some datasets have missing information without indications of it
 - Data errors
 - Data is sometimes of poor quality, erroneous, misspelled
 - Error propagation resulting from computer annotation

- Open access to sequences is not only essential for all of the work we do, if it was not there, there would be no bioinformatics, no BLAST, no CBP
- As critical as open access to sequence information is the open access to the literature.

What to take home

- Databases are a collection of data
 - Need to access and maintain easily and flexibly
- Biological information is vast and sometimes very redundant
- Computers can only create data, they do not give answers