# PR15_Genomske in proteomske podatkovne baze

# Biomolekularne podatkovne baze

# *Examples of biomolecular databases*

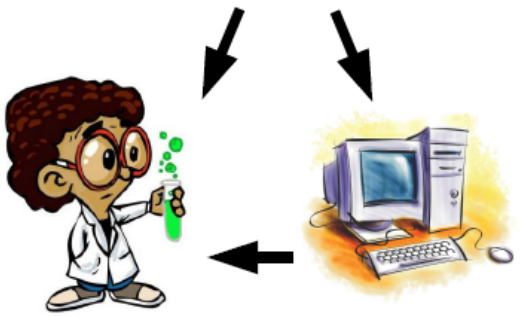
- Sequence and structure databases
  - Protein sequences (UniProt)
  - DNA sequences (EMBL, Genbank, DDBJ)
  - 3D structures (PDB)
  - Structural motifs (CATH)
  - Sequence motifs (PROSITE, PRODOM)
- Genome sequences and annotations
  - Genome-specific databases (SGD, FlyBase, AceDB, PlasmoDB, …)
  - 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

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

**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/

**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/

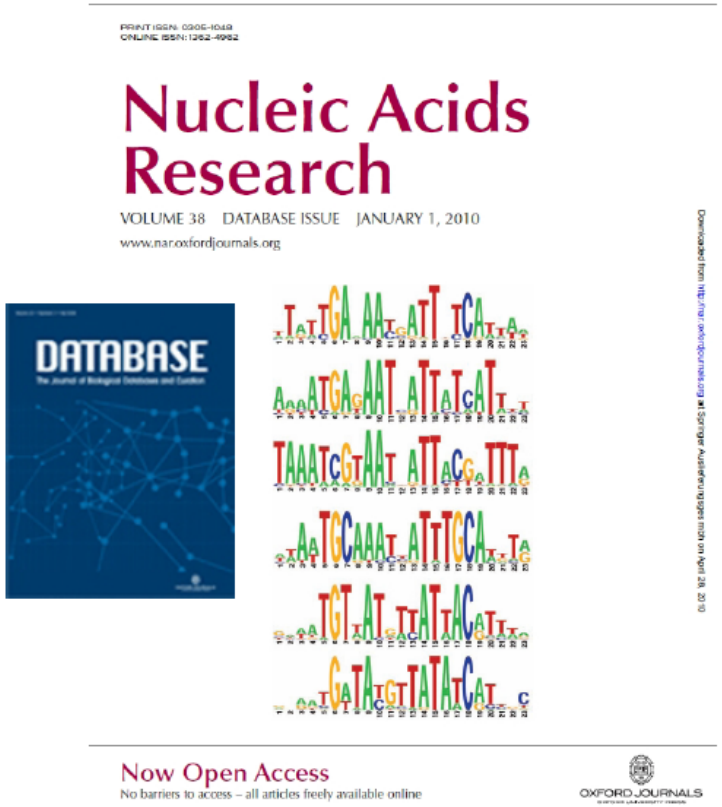
## Searching sequence databases - Secondary Databases

**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

**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.

# 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://lmgtfy.com/



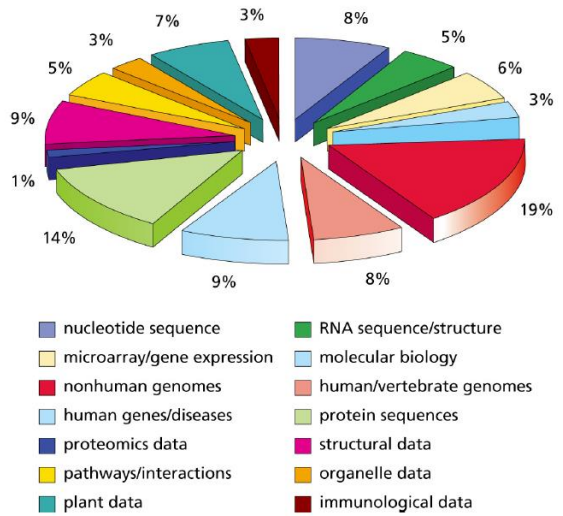
## *Databases of databases*

- 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/

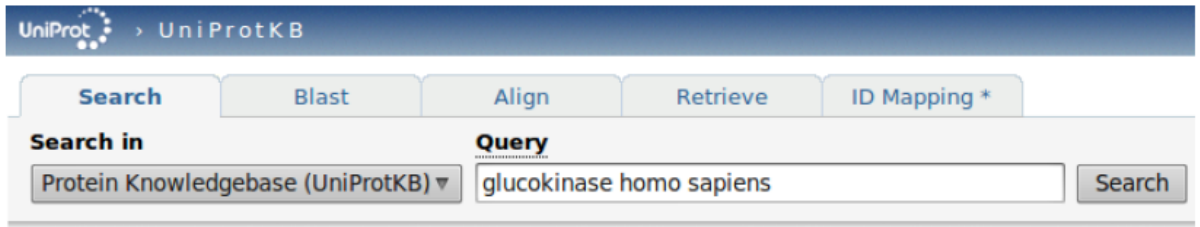


## 2014 NAR Database Summary Paper

- Compilation Paper
- Category List
- Alphabetical List
- Category/Paper List
- Search Summary Papers

# 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



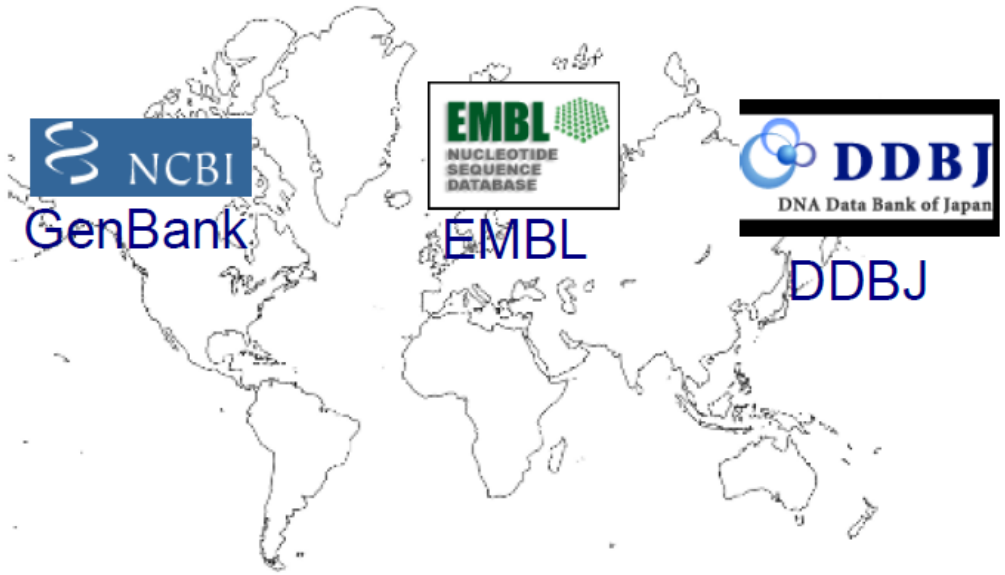

- **Web service** (SOAP, CORBA)
- **Flat files** (script based, large scale)
- **Database dump** (script based, large scale)

# 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

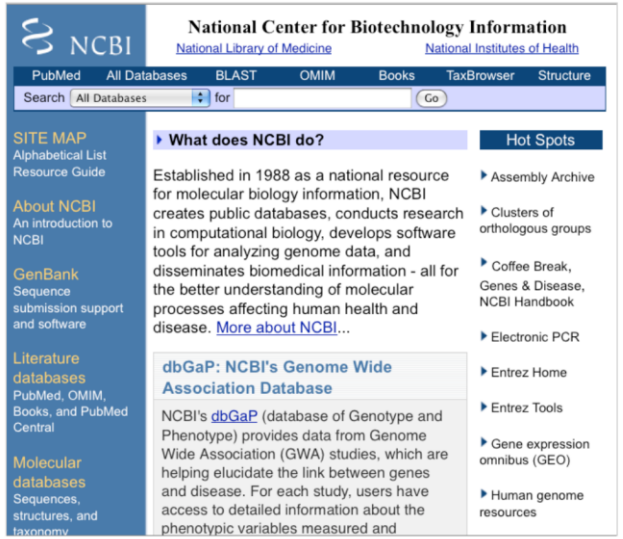



- sequences submitted directly by scientists and genome sequencing group, and sequences taken from literature and patents.
- entries in the EMBL, GenBank and DDBJ databases are **synchronized** on a daily basis
- **accession numbers** are managed in a consistent manner
- comparatively little error checking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

# Nucleic sequence databases: GenBank, EMBL, and DDBJ

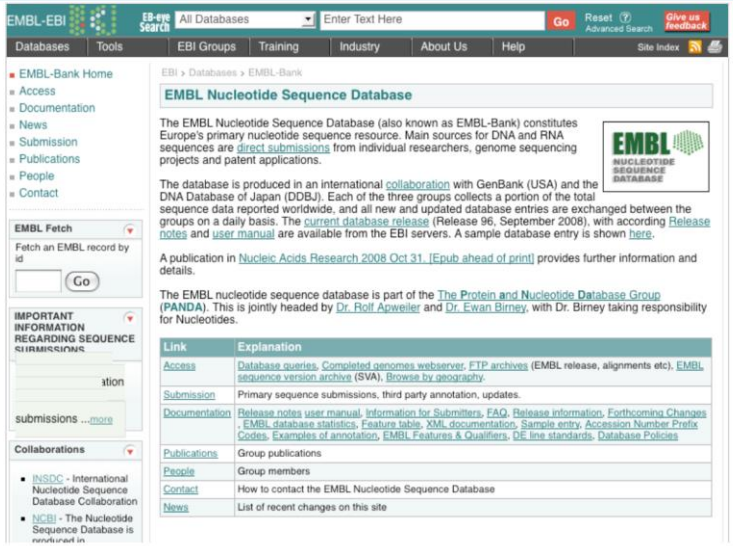
## Genbank (NCBI - USA)

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




## The EMBL Nucleotide Sequence Database (EBI - UK)

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




## DDBJ - DNA Data Bank of Japan

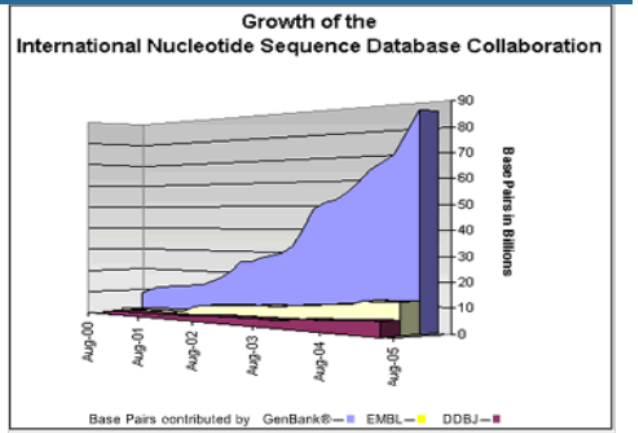
http://www.ddbj.nig.ac.jp/

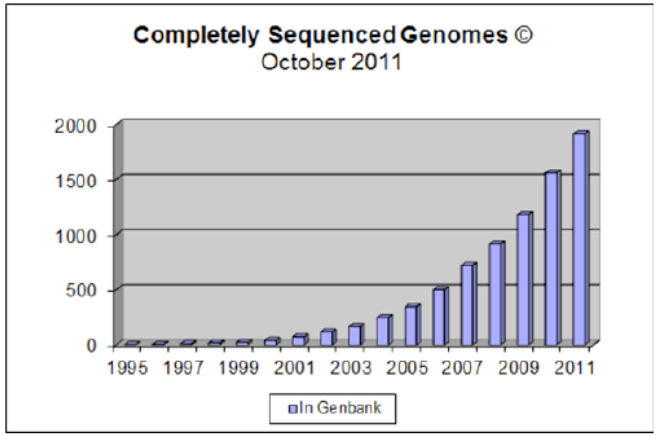

## Nucleic sequence databases

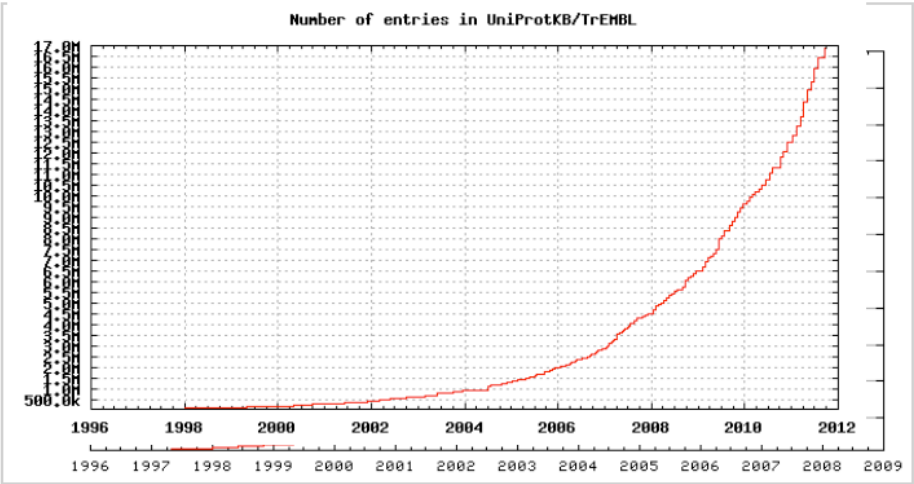
- To publish an article dealing with a sequence, scientific journals impose to have previously deposited this sequence in a reference database.
- There are 3 main repositories for nucleic acid sequences.
- Sequences deposited in any of these 3 databases are automatically synchronized in the 2 other ones.

# 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

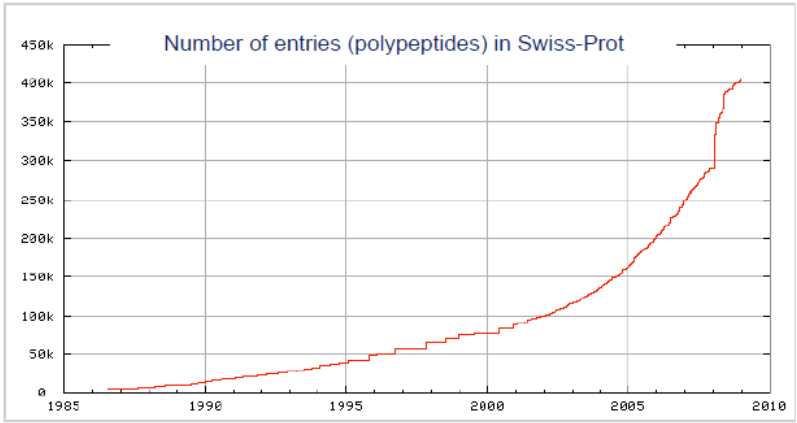
# Protein sequence databases

- 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**

# UniProt - the Universal Protein Resource

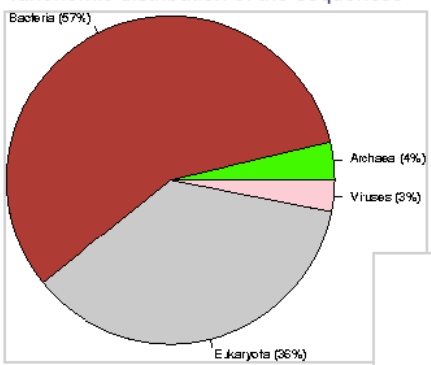
*http://www.uniprot.org/*

- 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.

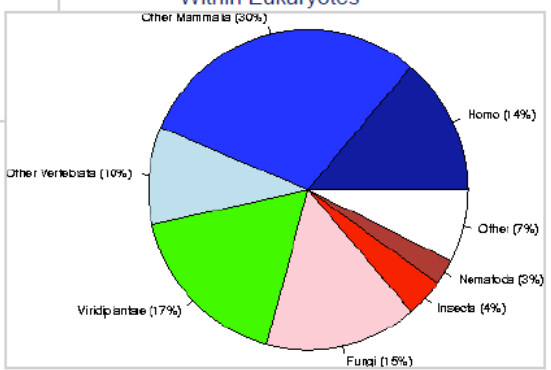



http://www.expasy.org/sprot/relnotes/relstat.html

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 | *Recommended name:*<br>**Paired box protein Pax-6**<br>*Alternative name(s):*<br>Oculorhombin<br>Aniridia type II protein |
| Gene names | Name: **PAX6**<br>Synonyms: AN2 |
| Organism | **Homo sapiens (Human)** |
| Taxonomic identifier | 9606 [NCBI] |
| Taxonomic lineage | Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › 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 | 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 prognostis 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.<br>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.<br>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.<br>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.<br>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.<br>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.<br>Defects in PAX6 are a cause of coloboma of optic nerve [MIM:120430].<br>Defects in PAX6 are a cause of bilateral optic nerve hypoplasia [MIM:165550]; also known as bilateral optic nerve aplasia. Inheritance is autosomal dominant. |
| Sequence similarities | Belongs to the paired homeobox family.<br>Contains 1 homeobox DNA-binding domain.<br>Contains 1 paired domain. |

## UniProt example - Human Pax-6 protein
## Structured annotation : keywords and Gene Ontology terms

**Ontologies** Hide | Top

**Keywords**

| | |
|---|---|
| Biological process | Differentiation<br>Transcription<br>Transcription regulation |
| Cellular component | Nucleus |
| Coding sequence diversity | Alternative splicing |
| Disease | Disease mutation |
| Domain | Homeobox<br>Paired box |
| Ligand | DNA-binding |
| Molecular function | Developmental protein<br>Repressor |
| Technical term | 3D-structure |

**Gene Ontology (GO)**

| | |
|---|---|
| Biological process | cell differentiation<br>Inferred from electronic annotation. Source: UniProtKB-KW<br>central nervous system development<br>Traceable author statement. Source: ProtInc<br>eye development<br>Traceable author statement. Source: ProtInc<br>organ morphogenesis [Ref.19]<br>Traceable author statement. Source: ProtInc<br>regulation of transcription, DNA-dependent<br>Inferred from electronic annotation. Source: InterPro<br>visual perception [Ref.19]<br>Traceable author statement. Source: ProtInc |
| Cellular component | nucleus<br>Inferred from electronic annotation. Source: InterPro |
| Molecular function | protein binding<br>Inferred from physical interaction. Source: IntAct<br>sequence-specific DNA binding<br>Inferred from electronic annotation. Source: InterPro<br>transcription factor activity [Ref.19]<br>Traceable author statement. Source: ProtInc |

Complete GO annotation...

## UniProt example - Human Pax-6 protein
## Protein interactions; Alternative products

**Binary interactions** Hide | Top

| With | Entry | #Exp. | IntAct | Notes |
|---|---|---|---|---|
| Dynll1 | 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** Hide | Top

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*;

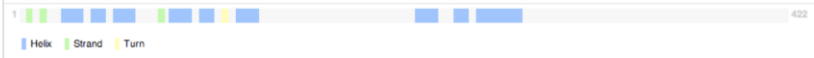
*The sequence of this isoform is not available.*

## UniProt example - Human Pax-6 protein
## Detailed description of regions, variations, and secondary structure

**Sequence annotation (Features)** Hide | Top

| | Feature key | Position(s) | Length | Description | Graphical view | Feature identifier |
|---|---|---|---|---|---|---|
| **Molecule processing** | | | | | | |
| | Chain | 1 – 422 | 422 | Paired box protein Pax-6 | | PRO_0000050185 |
| **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**




Details...

## UniProt example - Human Pax-6 protein
## Peptidic sequence

**Sequences** Hide | Top

| Sequence | | Length | Mass (Da) | Tools |
|---|---|---|---|---|
| **Isoform 1** [UniParc]. | FASTA | 422 | 46,683 | Blast go |

Last modified July 15, 1999. Version 2.
Checksum: C33CDD2C1B13C397

```
        10         20         30         40         50         60
MQNSHSGVNQ LGGVFVNGRP LPDSTRQKIV ELAHSGARPC DISRILQVSN GCVSKILGRY

        70         80         90        100        110        120
YETGSIRPRA IGGSKPRVAT PEVVSKIAQY KRECPSIFAW EIRDRLLSEG VCTNDNIPSV

       130        140        150        160        170        180
SSINRVLRNL ASEKQQMGAD GMYDKLRMLN GQTGSWGTRP GWYPGTSVPG QPTQDGCQQQ

       190        200        210        220        230        240
EGGGENTNSI SSNGEDSDEA QMRLQLKRKL QRNRTSFTQE QIEALEKEFE RTHYPDVFAR

       250        260        270        280        290        300
ERLAAKIDLP EARIQVWFSN RRAKWRREEK LRNQRRQASN TPSHIPISSS FSTSVYQPIP

       310        320        330        340        350        360
QPTTPVSSFT SGSMLGRTDT ALTNTYSALP PMPSFTMANN LPMQPPVPSQ TSSYSCMLPT

       370        380        390        400        410        420
SPSVNGRSYD TYTPPHMQTH MNSQPMGTSG TTSTGLISPG VSVPVQVPGS EPDMSQYWPR


LQ
```

« Hide

| | | | | |
|---|---|---|---|---|
| **Isoform 5a** (Pax6-5a) [UniParc]. | | 436 | 48,218 | Blast go |

Checksum: 74926827347A20B5
Show »

**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] **"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.
Cell 67:1059-1074(1991) [PubMed: 1684738] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [MRNA].

[2] **"Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene."**
Glaser T., Walton D.S., Maas R.L.
Nat. Genet. 2:232-239(1992) [PubMed: 1345175] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [MRNA].

[3] Liu J., Zhang B., Zhou Y., Peng X., Yuan J., Qiang B.
Submitted (JUL-2001) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE (ISOFORM PAX6).

[4] The German cDNA consortium
Submitted (AUG-2003) to the EMBL/GenBank/DDBJ databases
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORM 5A)

[24] **"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.
Graefes Arch. Clin. Exp. Ophthalmol. 238:552-558(2000) [PubMed: 10955655] [Abstract]
Cited for: VARIANT NYSTAGMUS ARG-118.

[25] **"Missense mutation at the C-terminus of PAX6 negatively modulates homeodomain function."**
Singh S., Chao L.-Y., Mishra R., Davies J., Saunders G.F.
Hum. Mol. Genet. 10:911-918(2001) [PubMed: 11309364] [Abstract]
Cited for: VARIANTS AN2 GLN-375 AND ARG-422.

[26] **"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.
Am. J. Hum. Genet. 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.

\+ Additional computationally mapped references.

## *UniProt example - Human Pax-6 protein*
## *Cross-references to many databases (fragment shown)*

**Sequence databases**

| | |
|---|---|
| EMBL | M77844 mRNA. Translation: AAA59963.1.<br>M77844 mRNA. Translation: AAA59962.1.<br>M93650 mRNA. Translation: AAA36416.1.<br>AY047583 mRNA. Translation: AAK95849.1.<br>BX640762 mRNA. Translation: CAE45868.1.<br>Z95332, Z83307 Genomic DNA. Translation: CAG38363.1.<br>Z83307, Z95332 Genomic DNA. Translation: CAG38087.1.<br>BC011953 mRNA. Translation: AAH11953.1. |
| PIR | A56674. |
| RefSeq | NP_000271.1.<br>NP_001121084.1.<br>NP_001595.2. |
| UniGene | Hs.591993 |

**3D structure databases**

PDB

| 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. |

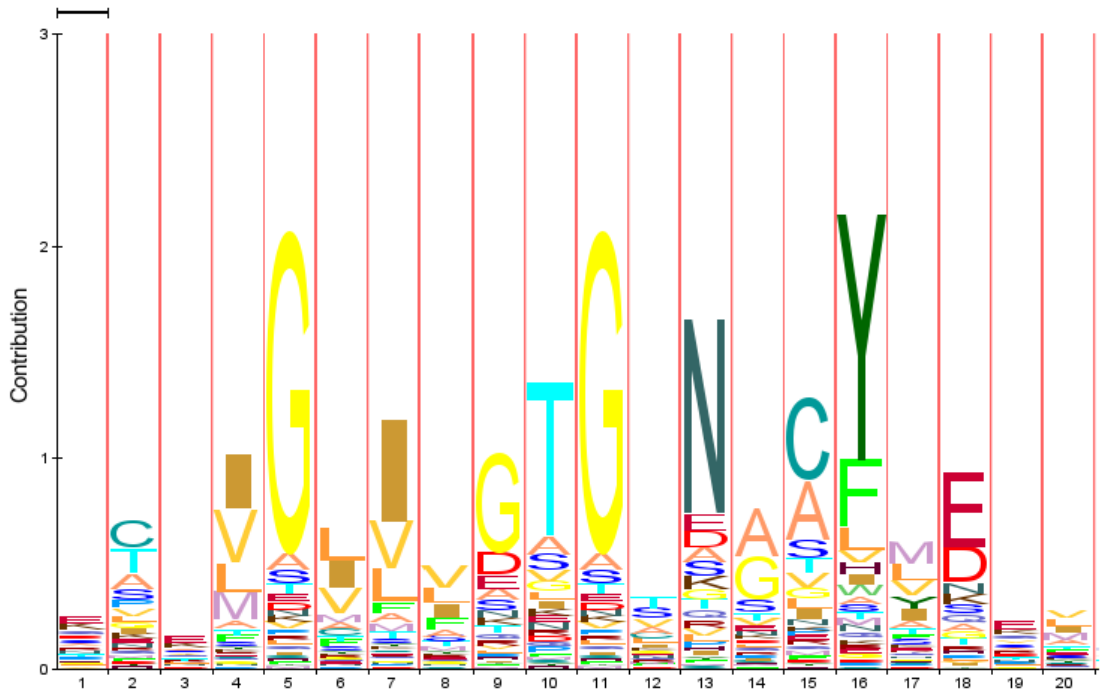
**Genome annotation databases**

| | |
|---|---|
| Ensembl | ENSG00000007372. Homo sapiens. [Contig view] |
| GeneID | 5080. |
| KEGG | hsa:5080. |

# Sekundarne biološke podatkovne baze

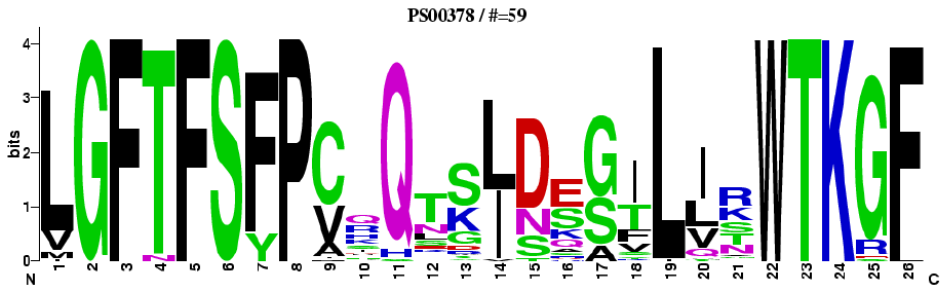
# Secondary Databases

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



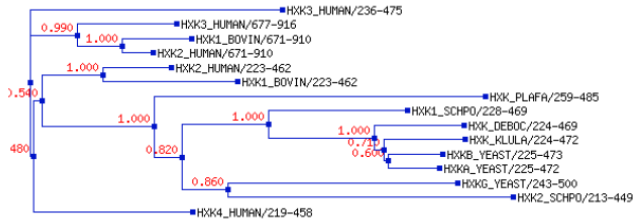
# Motifs and secondary structure

- 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<br>View Pfam proteome data. |
| **Length:** | 917 amino acids |

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

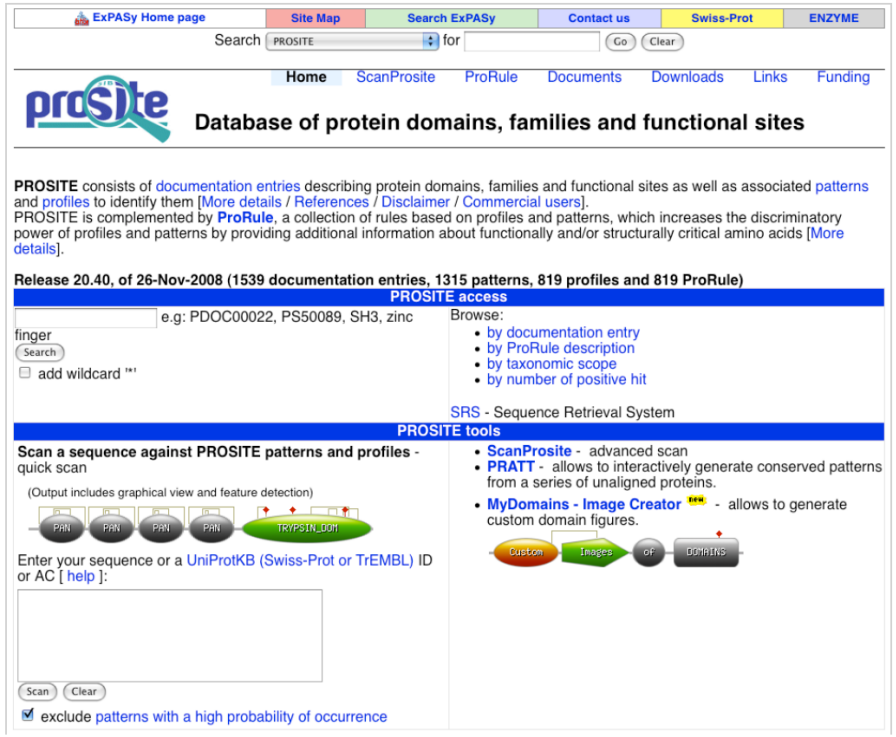
**Pfam domains**

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 |

# Prosite - protein domains, families and functional sites

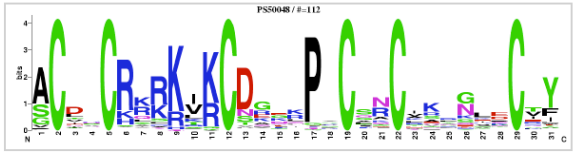
*http://www.expasy.ch/prosite/*

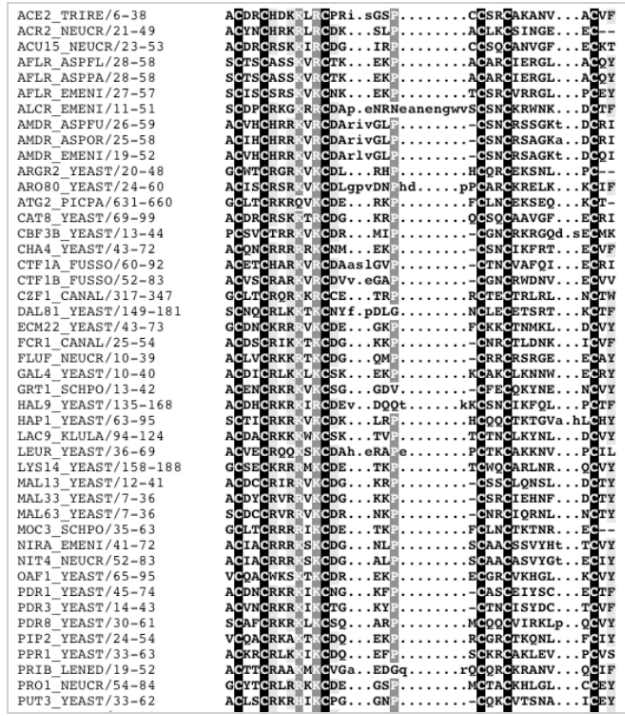



# Prosite - aligned sequences and logo

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

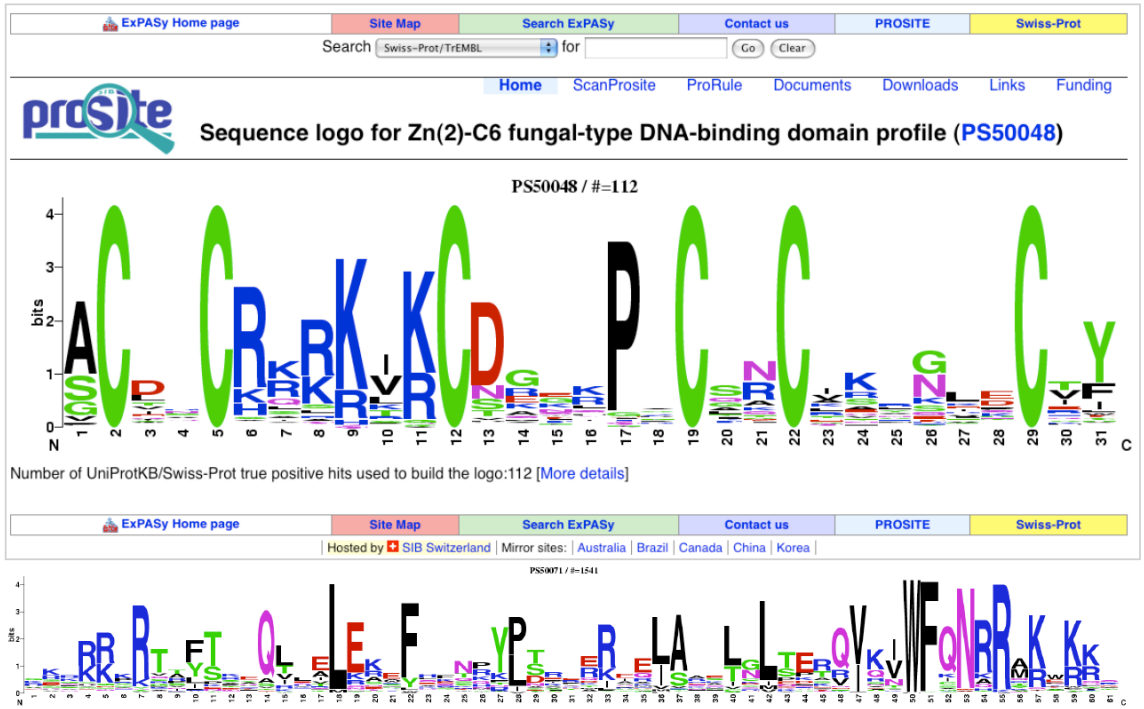
- Some of the sequences that were used to built the Prosite profile for the Zn(2)-C6 fungal-type DNA-binding domain (ZN2_CY6_FUNGAL_2, PS50048).
- The Sequence Logo (below) indicates the level of conservation of each residue in each column of the alignment.
- Note the 6 cysteines, characteristic of this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



# Prosite - Example of sequence logo

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







# Prosite - Example of domain signature

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

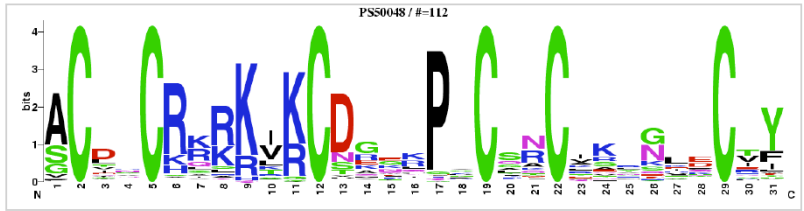
- The domain signature is a string-based pattern representing the residues that are characteristic of a domain.

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. |

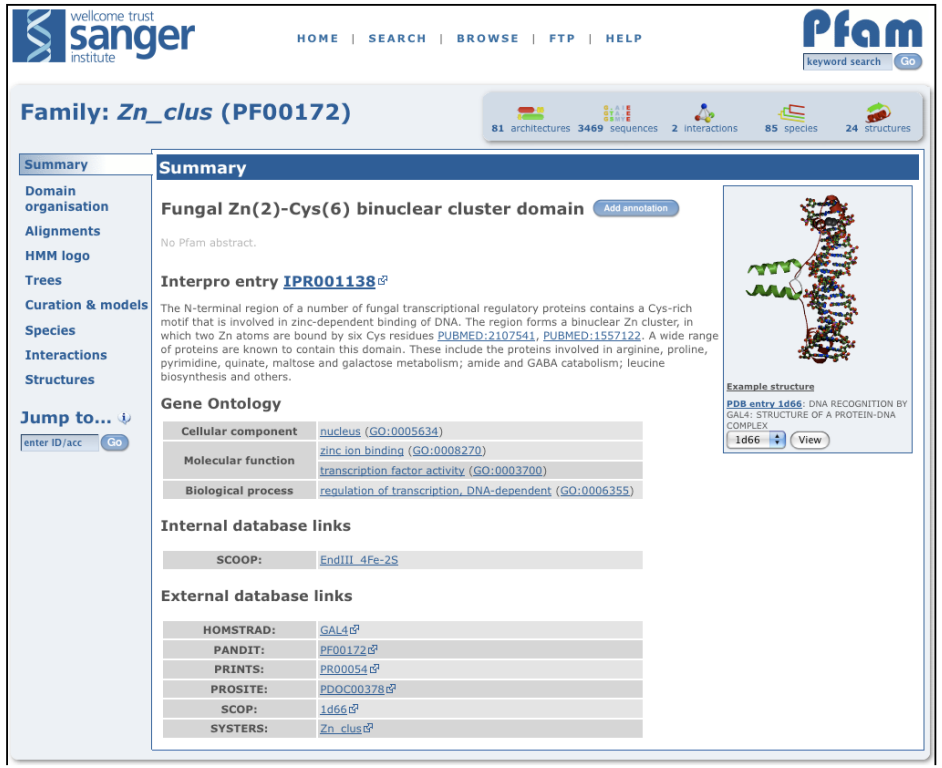
- Retrieve an alignment of Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS00463
- Retrieve a list 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]

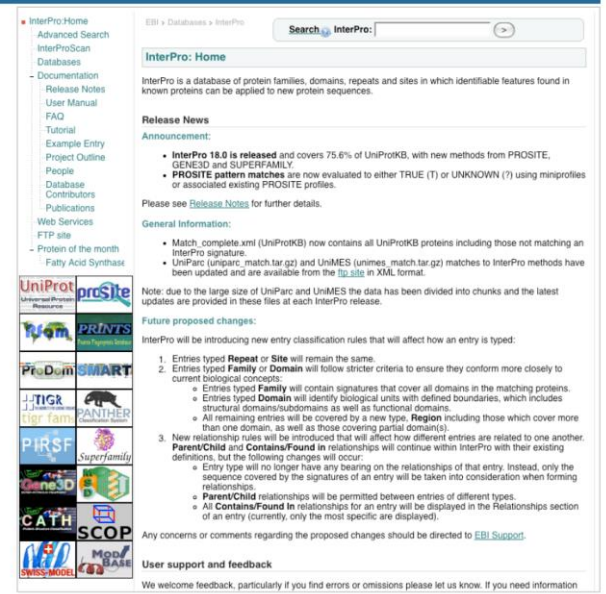


## PFAM (Sanger Institute - UK) http://pfam.sanger.ac.uk/

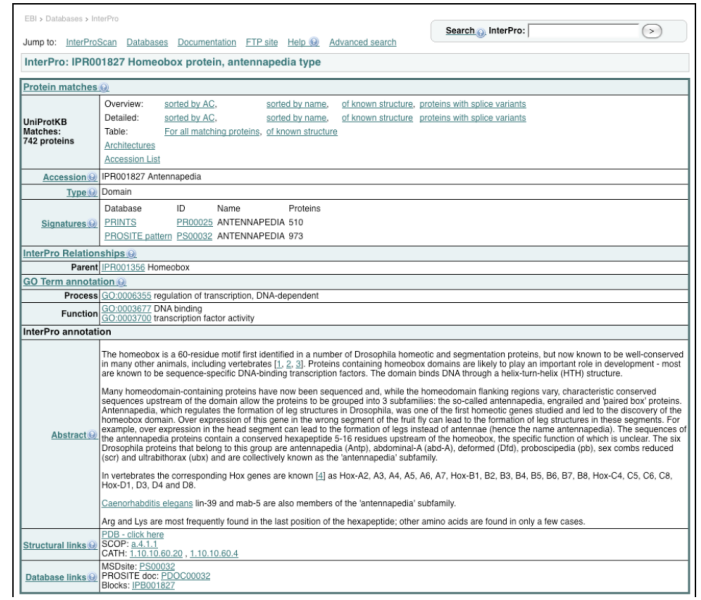
*Protein families represented by multiple sequence alignments and hidden Markov models (HMMs)*




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




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



# Peptide related information

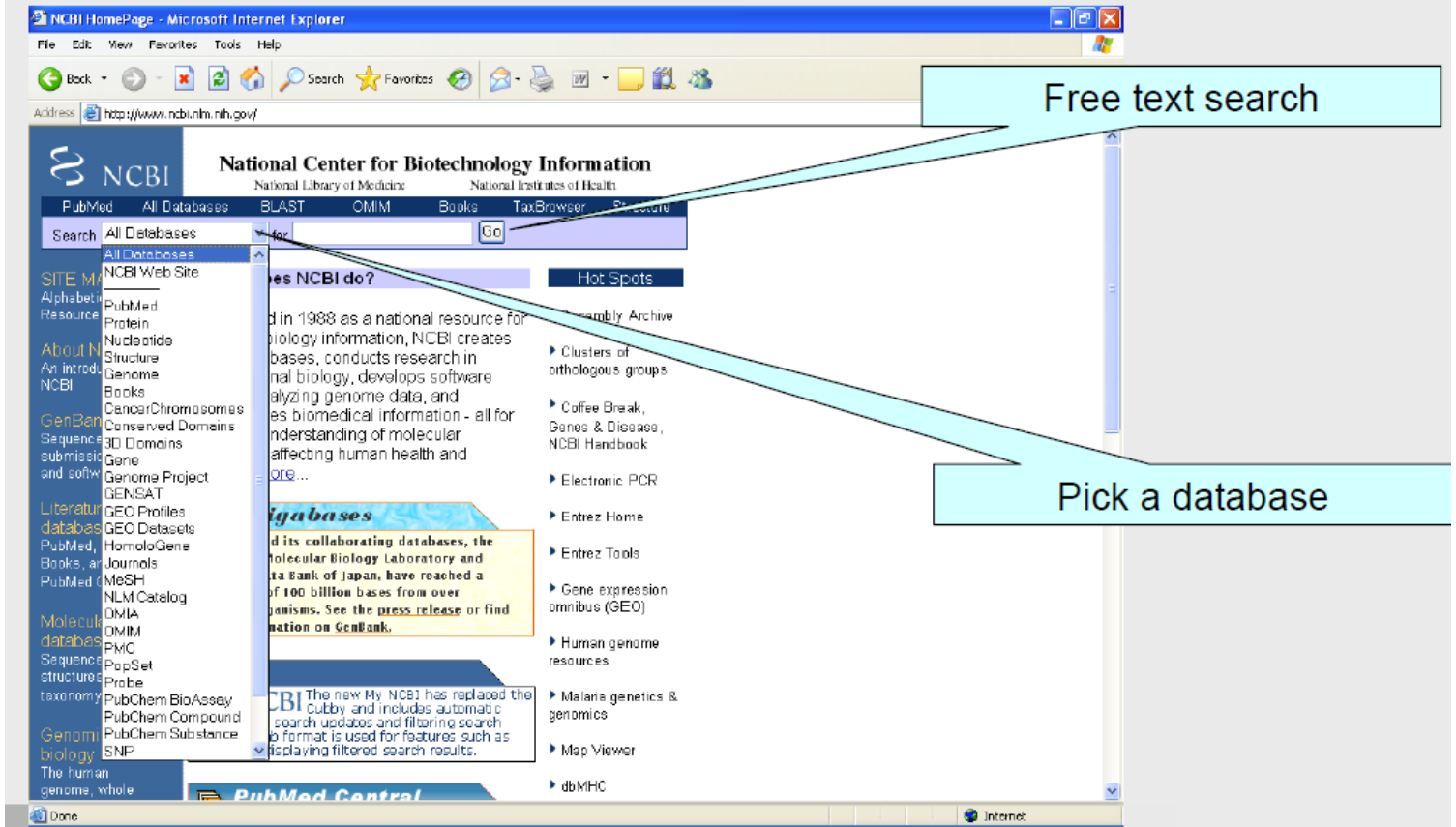
- MEROPS - Peptidase Database
- Peptide Database (Cancer) [example]
- PeptideMass
  - cleaves a protein sequence from the UniProt Knowledgebase (Swiss-Prot and TrEMBL) or a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides.
- SYSFPEITHI
  - SYFPEITHI 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

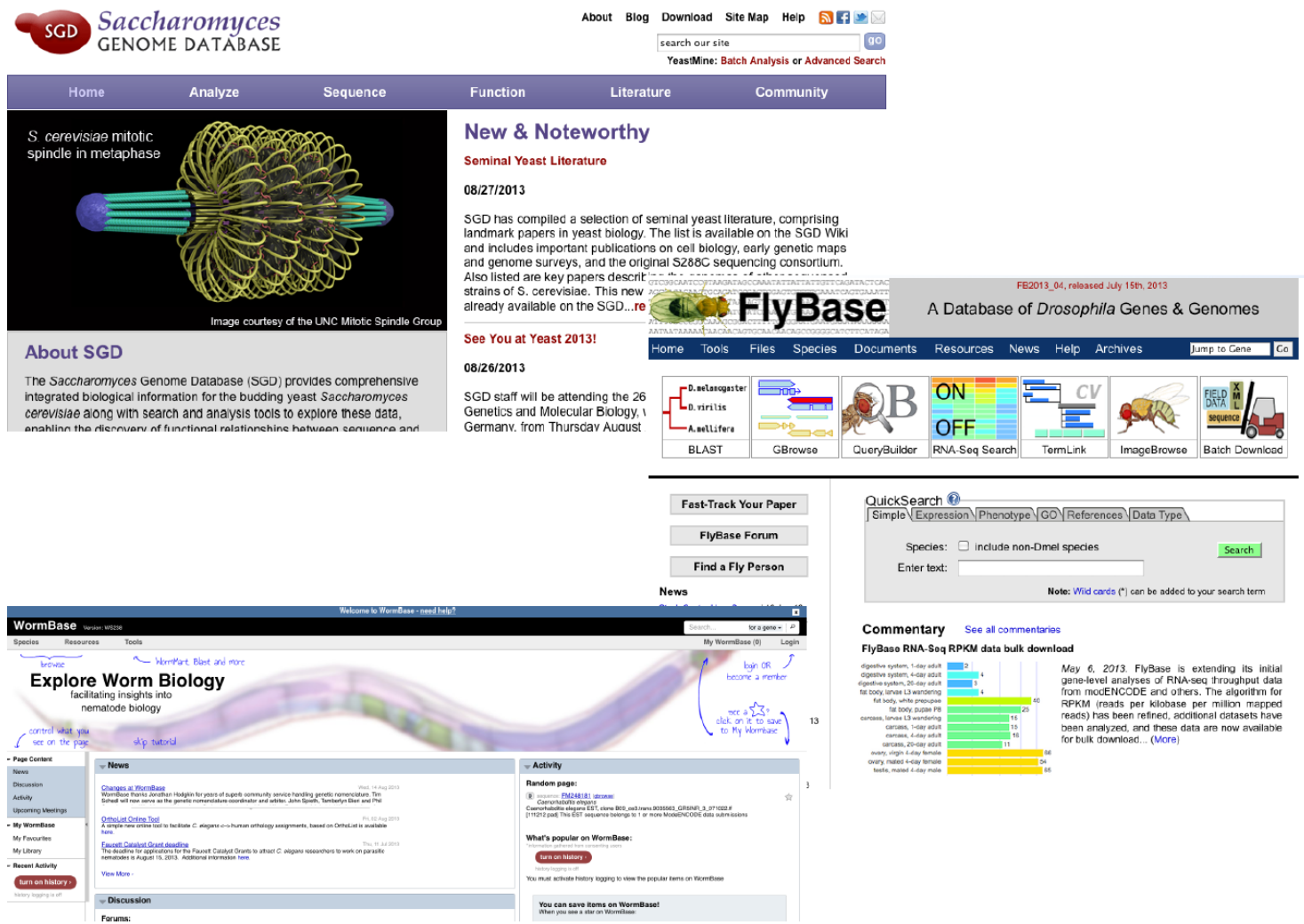
These databases of databases collect data from different sources and make them available in new and more convenient form, or with an emphasis on a particular disease or organism.

# Entrez – main page

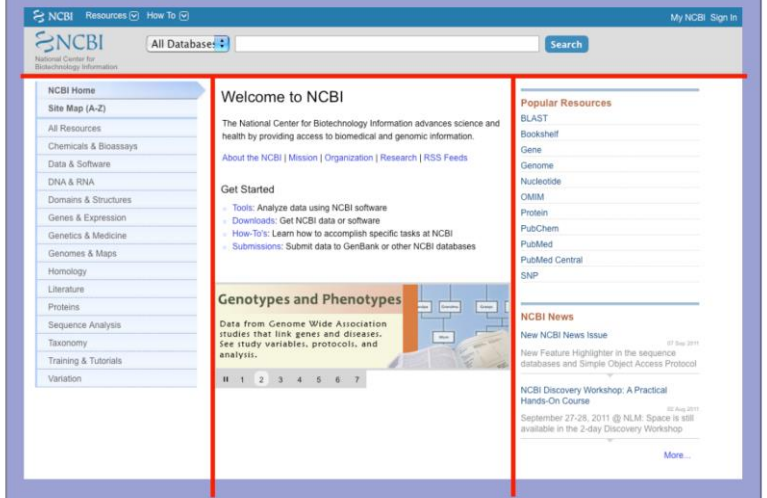


# Many Datasets at NCBI

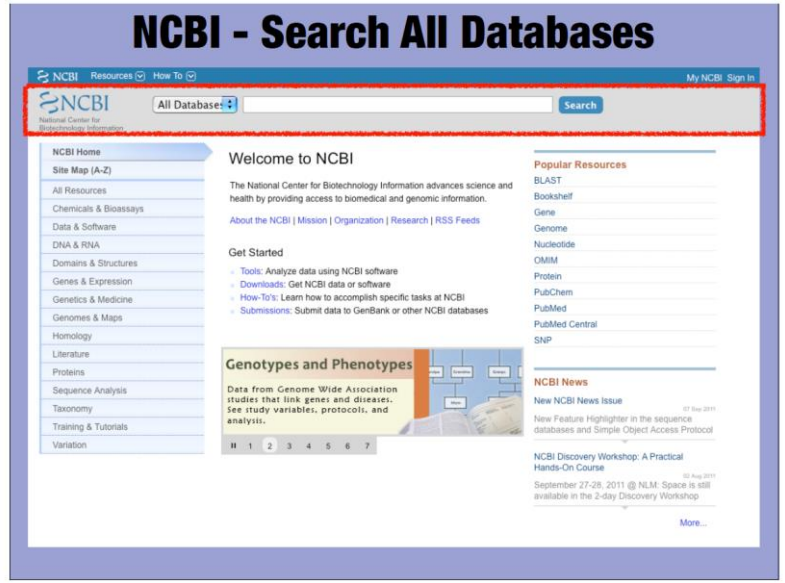
- The NCBI hosts a huge interconnected database system that, in addition to DNA and protein, includes:
  - Journal Articles (PubMed)
  - Genetic Diseases (OMIM)
  - Polymorphisms (dbSNP)
  - Cytogenetics (CGH/SKY/FISH & CGAP)
  - Gene Expression (GEO)
  - Taxonomy
  - Chemistry (PubChem)



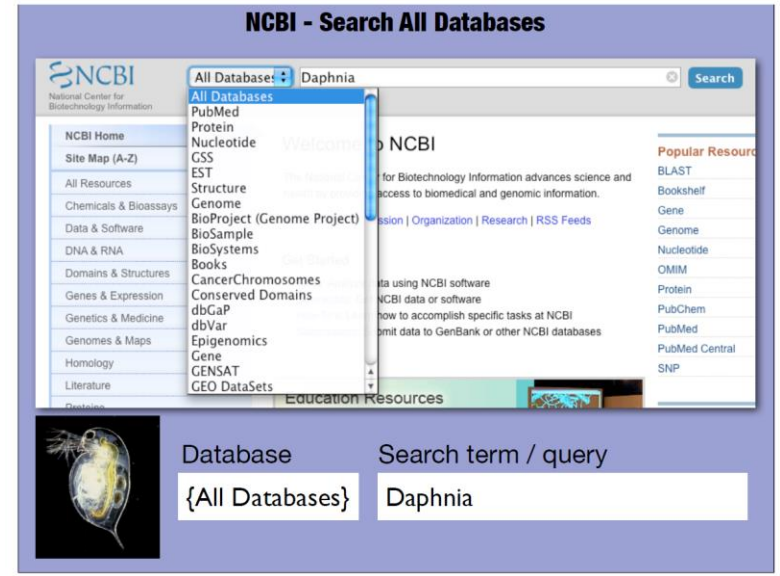
# NCBI - Welcome Website




# NCBI - Search All Databases

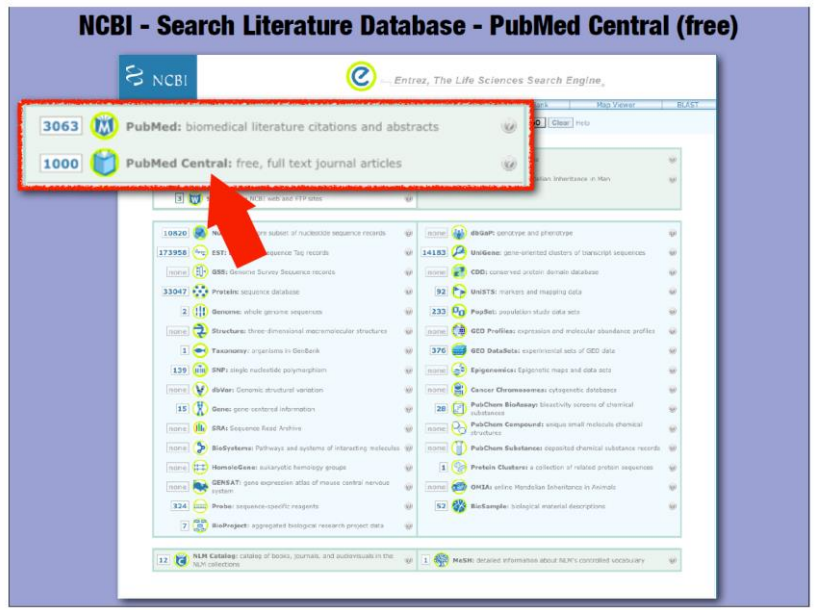



# NCBI - Search All Databases




# 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

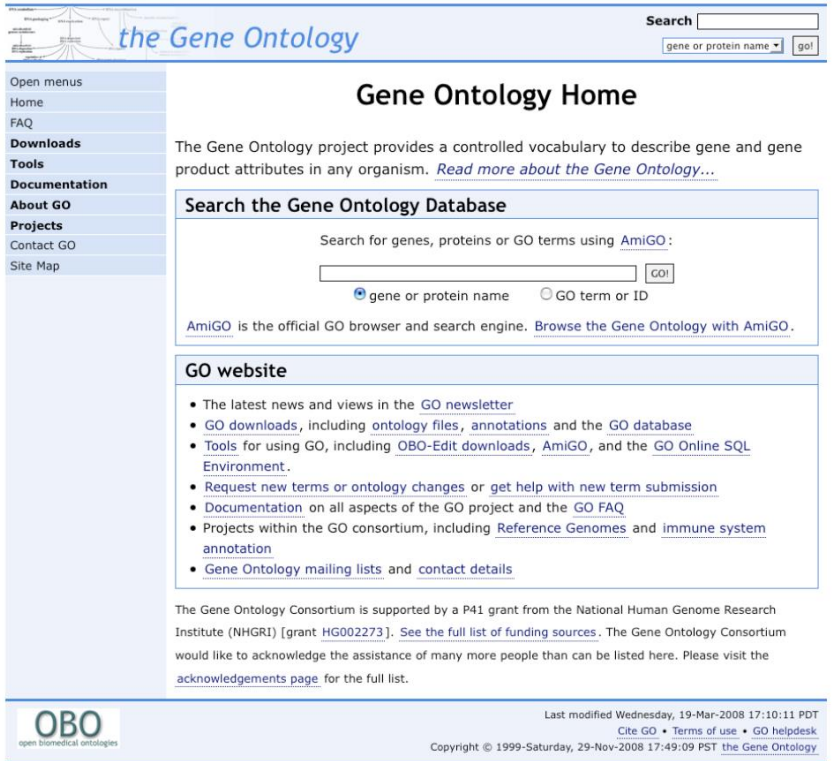
BioMart is a query-oriented data management system developed jointly by the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI).
The system can be used with any type of data and is particularly suited for providing 'data mining' like searches of complex descriptive data.

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

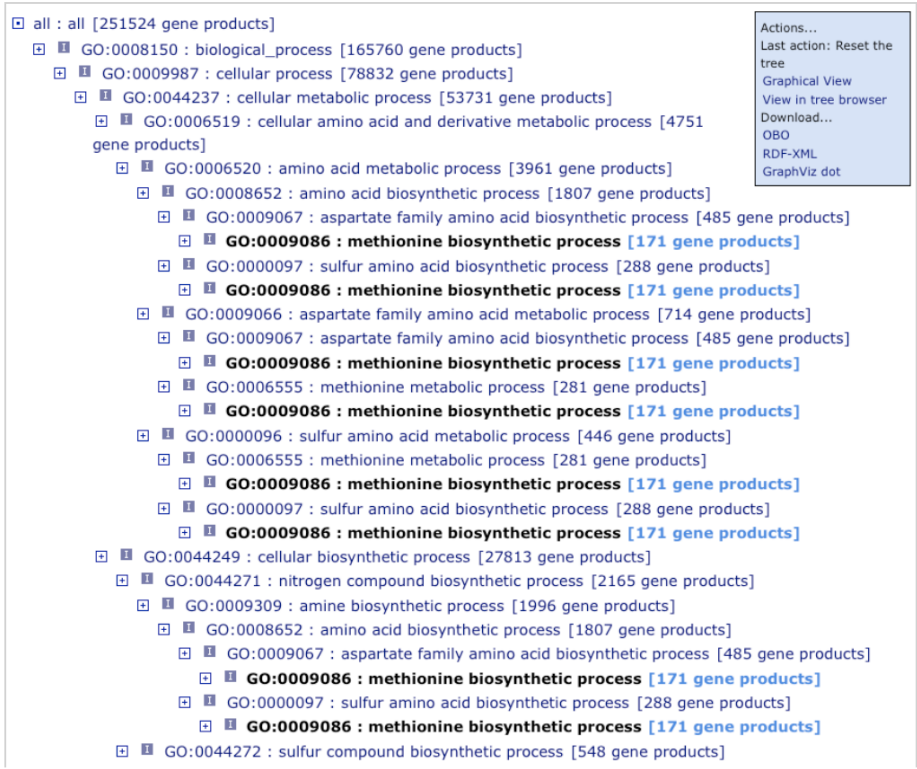
Web services are application programming interfaces (API) or web APIs that are accessed via Hypertext Transfer Protocol and executed on a remote system hosting the requested services

# Specializirane biološke podatkovne baze

# Gene Ontology Database
http://www.geneontology.org/

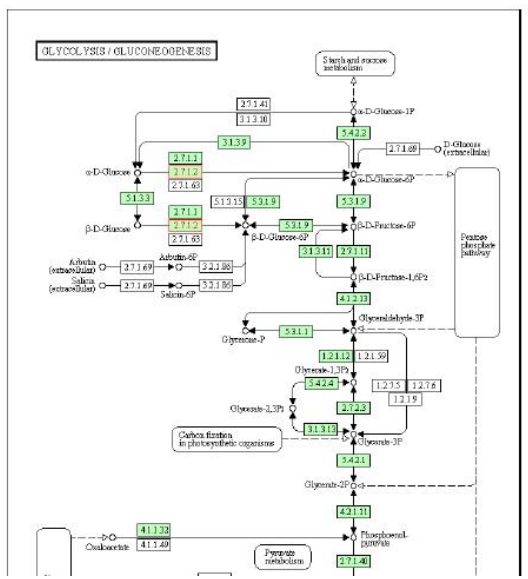


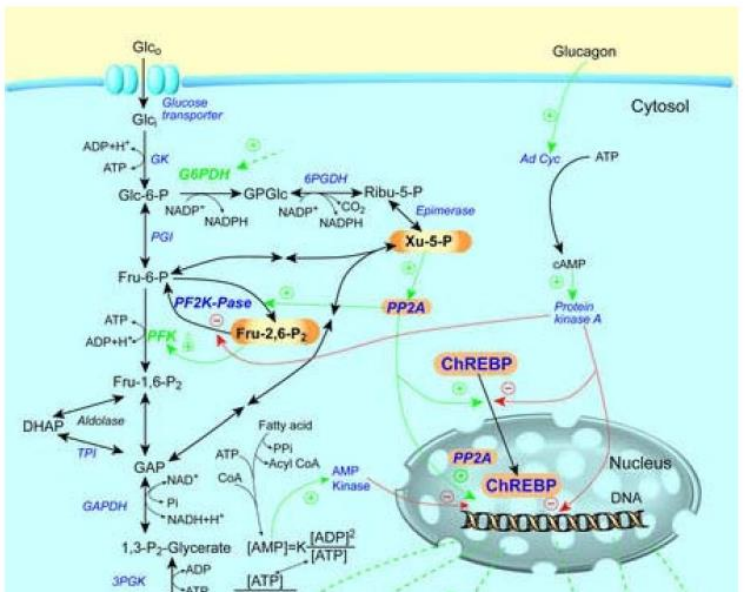

# Gene Ontology Database (http://www.geneontology.org/)
# Example: methionine biosynthetic process



# Metabolic networks - pathways

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

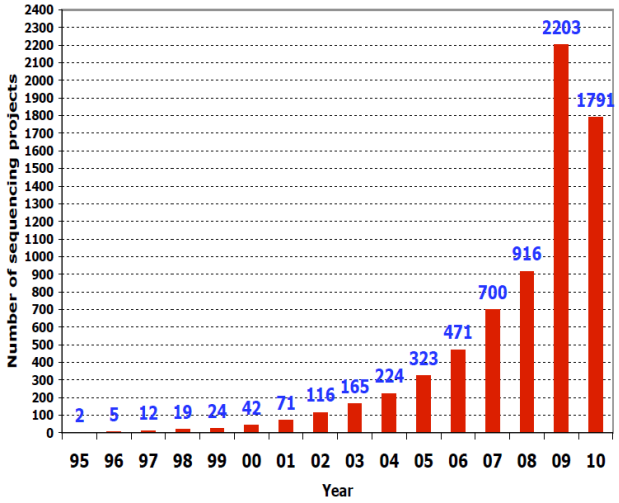


# Genomske podatkovne baze

# 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**

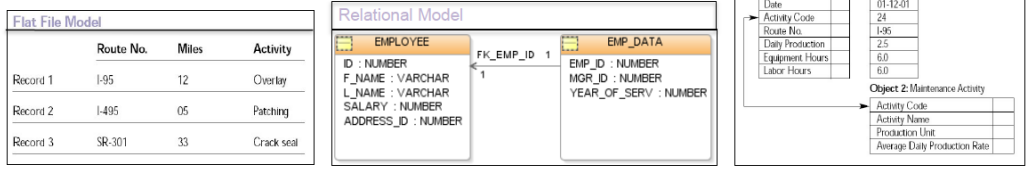



**Number of sequencing projects,** 1995-2010
**Source**: GOLD www.genomesonline.org

# 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:**



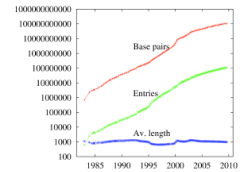

*Structured Files* *Tables* *Objects*

# How big are genomes ?

Top 20 organisms in public banks

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 | 2,556,492 |
| 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 |




Center for Biological Sequence Analysis
The Technical University of Denmark

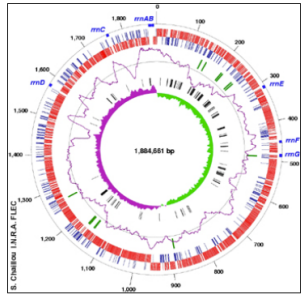
**DOGS – Databases Of Genome Sizes**
www.cbs.dtu.dk/databases/DOGS

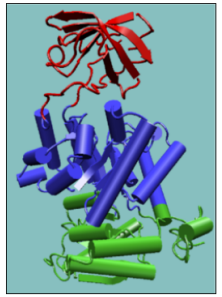
**Genome typical sizes:**

- 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...



```
>P03004-1
MSLSLWQQCLARLQDELPATEFSMWIRPLQAELSDNTLALYAPNRFVLDWVRDKYLNNIN
GLLTSFCGADAPQLRFEVGTKPVTQTPQAAVTSNVAAPAQVAQTQPQRAAPSTRSGWDNV
PAPAEPTYRSNVNVKHTFDNFVEGKSNQLARAAARQVADNPGGAYNPLFLYGGTGLGKTH
LLHAVGNGIMARKPNAKVVYMHSERFVQDMVKALQNNAIEEFKRYYRSVDALLIDDIQFF
ANKERSQEEFFHTFNALLEGNQQIILTSDRYPKEINGVEDRLKSRFGWGLTVAIEPPELE
TRVAILMKKADENDIRLPGEVAFFIAKRLRSNVRELEGALNRVIANANFTGRAITIDFVR
EALRDLLALQEKLVTIDNIQKTVAEYYKIKVADLLSKRRSRSVARPRQMAMALAKELTNH
SLPEIGDAFGGRDHTTVLHACRKIEQLREESHDIKEDFSNLIRTLSS
```

```
ID   AE001437_2; parent: AE001437
AC   AE001437; AE007513-AE007868;
FT   gene            467..1807
FT                   /gene="dnaA"
FT                   /locus_tag="CA_C0001"
SQ   Sequence  1341 BP;
     atggctgccc aattaaatga actgtggcaa aaaaccataa acataataaa aggtgaatta        60
     acggaagtaa gctttaacac ttggataaaa agtatcactc ctatatctat tgataaagat       120
     agtatacgcc ttagtgttcc aaatcaattt acaaaggaaa ttctcgaaaa tcgataataa       180
     gatttaataa taaattctat gaaaataata actactaaaa aatatgacat agcatttta        240
     atttcctctg aagaagcttt ggaaaccgat gaagatcaag aaaccgatac aaacaatgta       300
     aatacagata ctagttcttc tatgcttaat ccaaaataca aatttgattc atttgtaatt       360
     ggtaatagta atagatttgc tcatgctgca tgcctgctg ttgctgaagc tcctgcaaaa        420
     gcctacaatc cattatttat atatggagga gttggacttg gaaagactca tcttatgcaa       480
     gctataggtc attacatact tgataataat ccaaaagcta aagttgtata tgtttcttct       540
```

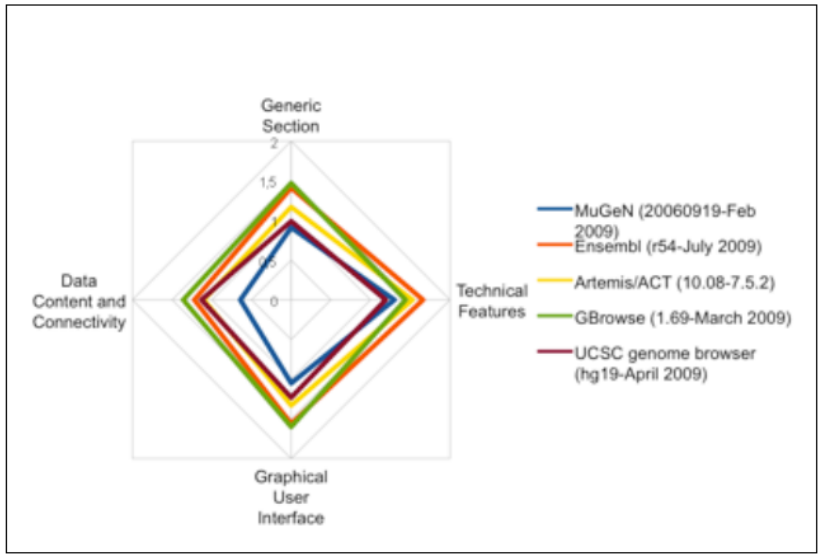
# Data visualization

**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

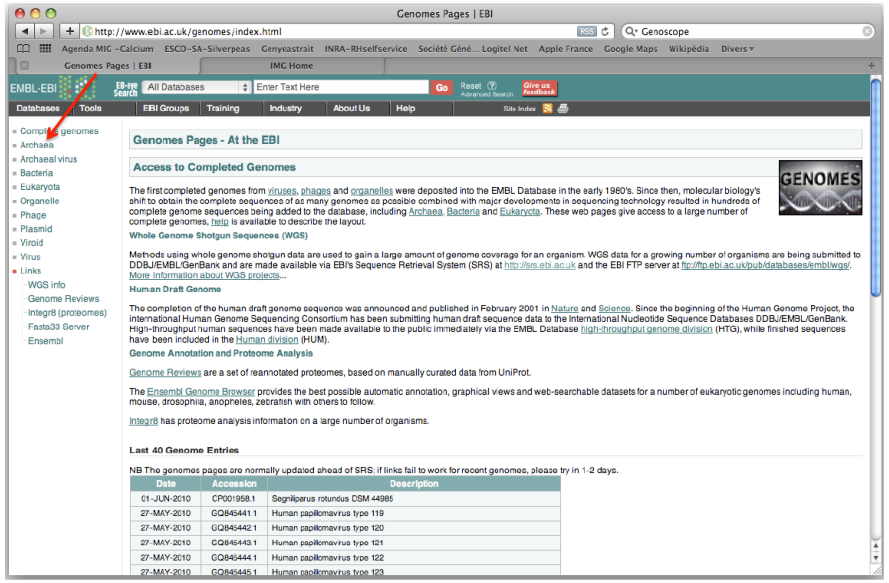


# Main resources for genomes

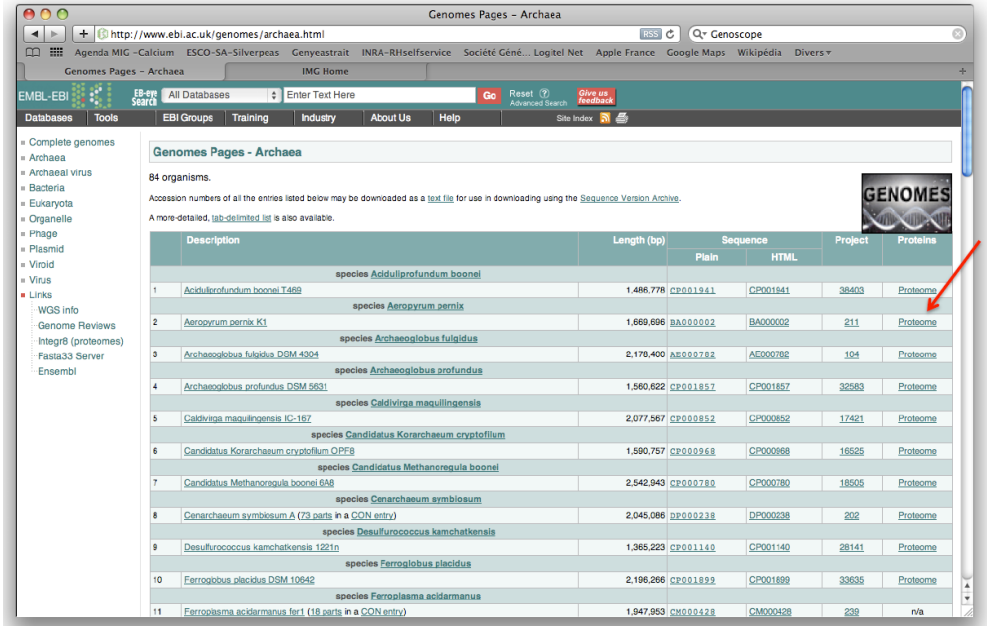
- Two main international resources:
  - **EBI: European Bioinformatics Institute**
    www.ebi.ac.uk/genomes
  - **NCBI: National Center for Biotechnology Information**
    www.ncbi.nlm.nih.gov/Genomes

- Many other resources from sequencing institutions:
  - **Sanger:** the Welcome 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 (Havard, MIT) — www.broadinstitute.org
  - **Genoscope:** the french sequencing center — www.genoscope.cns.fr
  - ….

# Browsing genomes at EBI

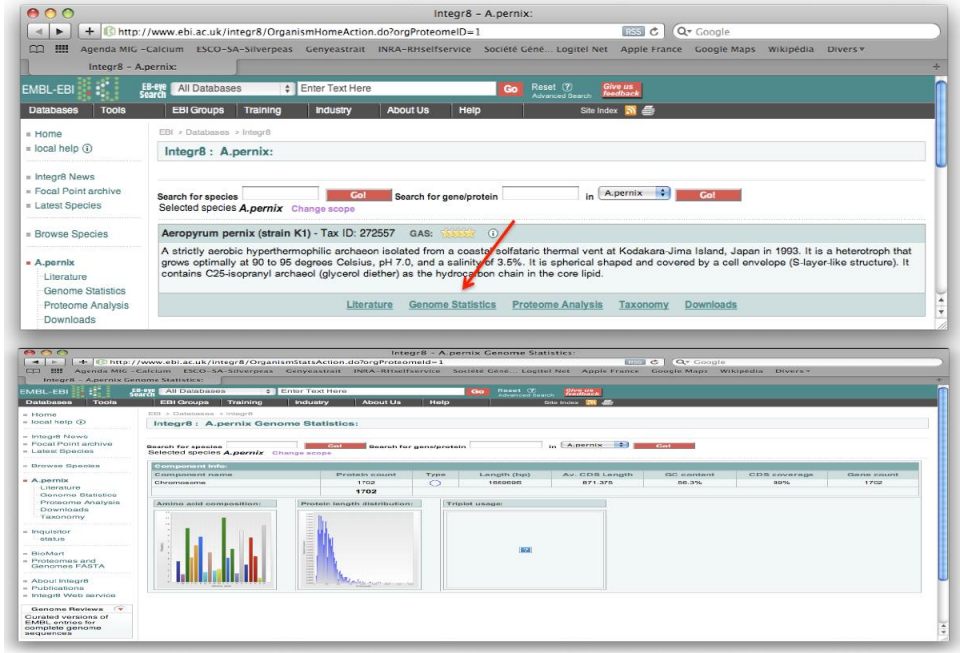



# EBI: genomes of archaea

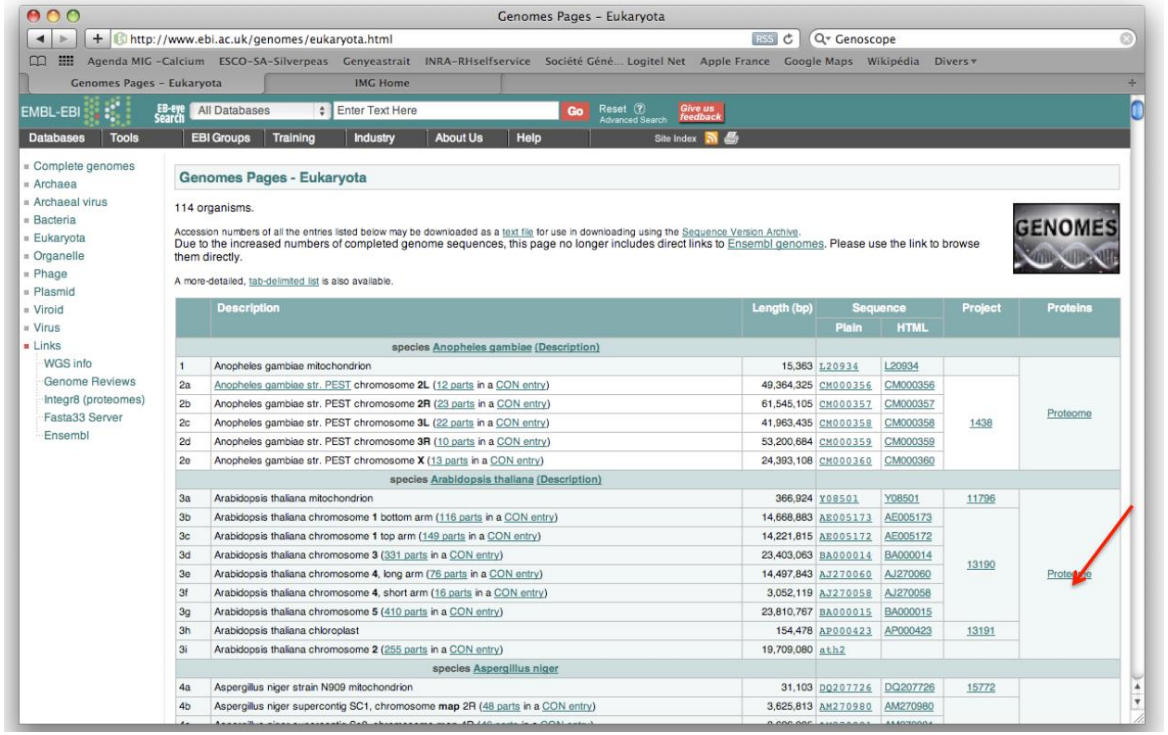



Genome databases

# EBI: genome of A. pernix K1

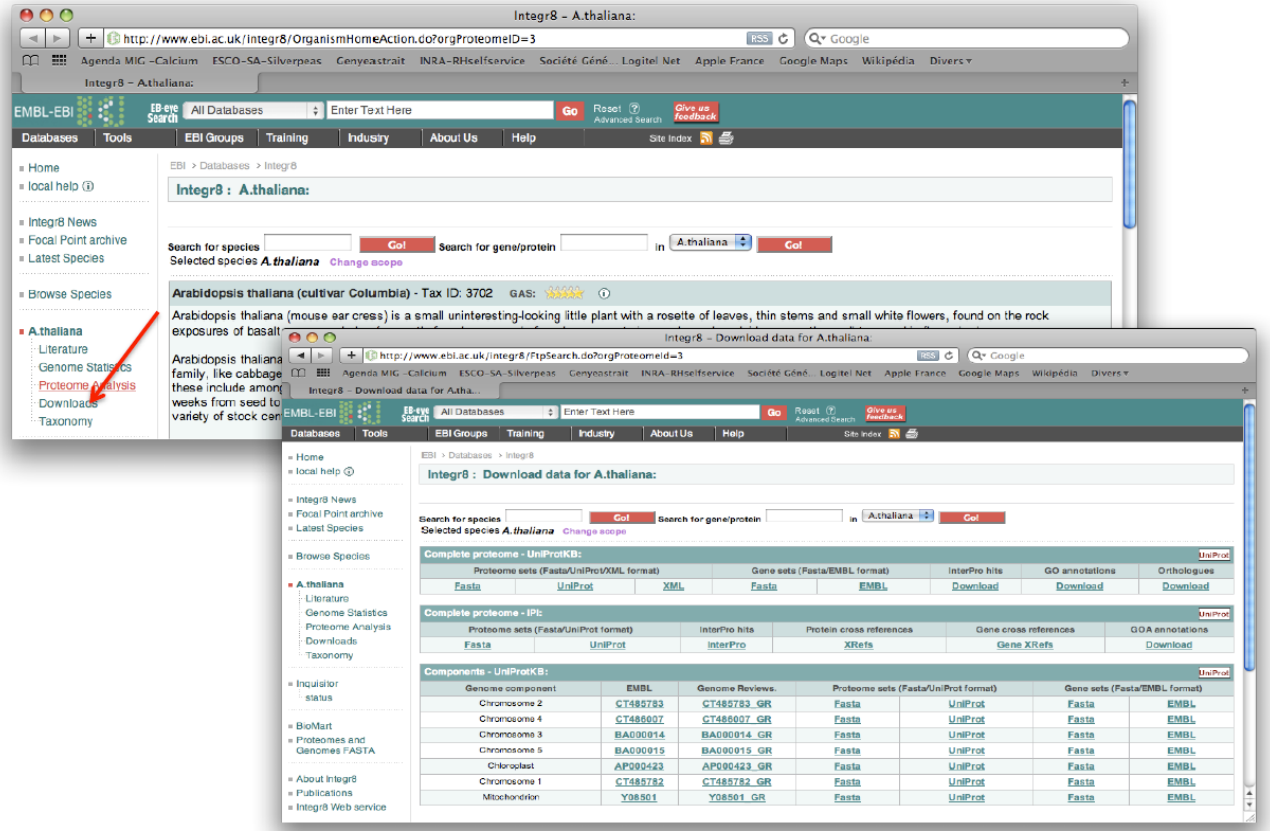


# EBI: genomes of eukaryotes

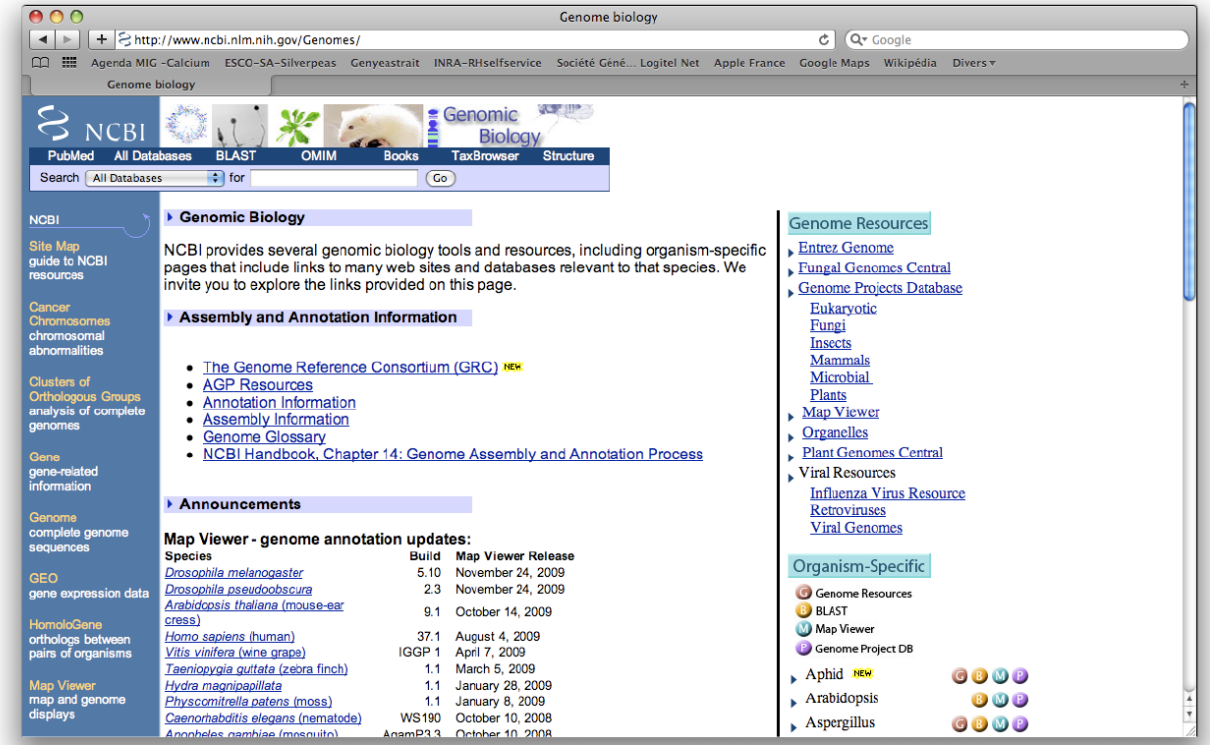





# EBI: download genomic data of *A. thaliana*

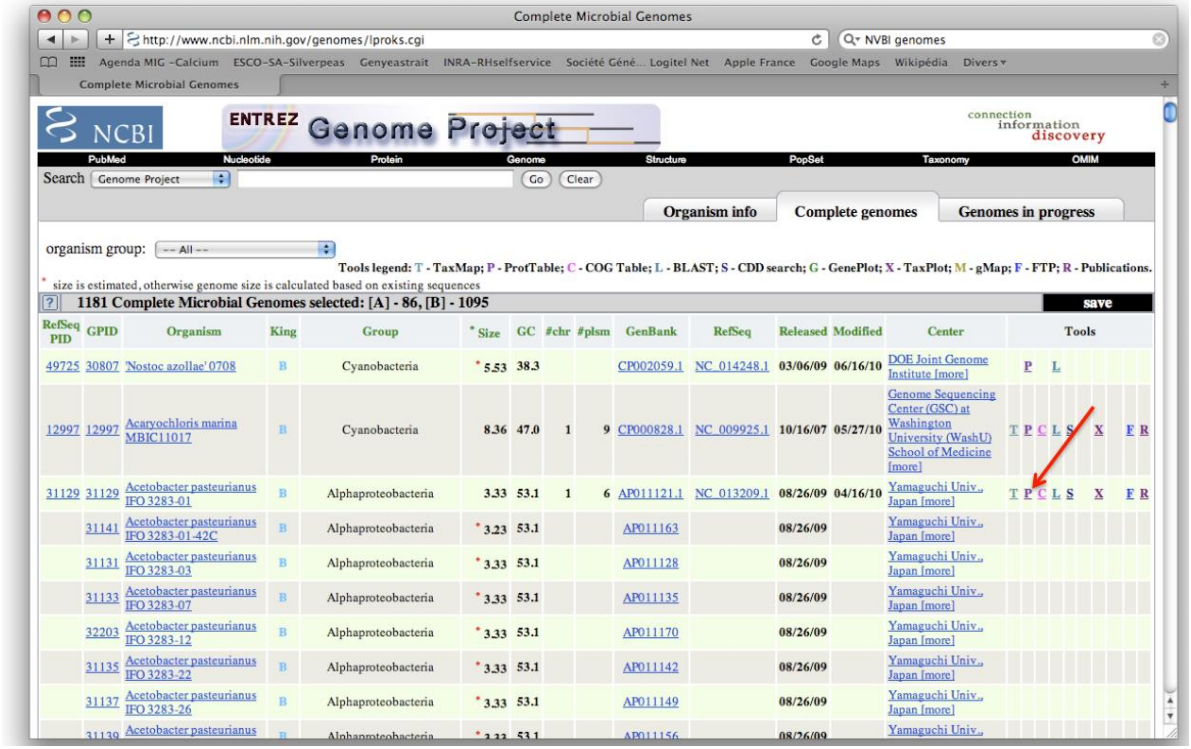


# Browsing genomes at NCBI

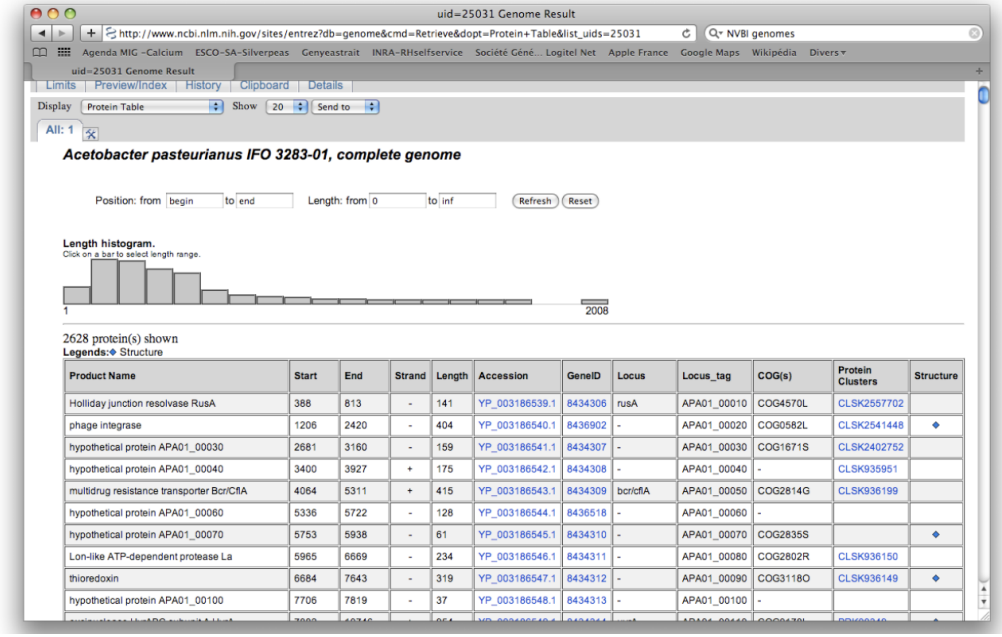





# Prokaryotes genomes at NCBI



## *Acetobacter Pasteurianus* proteome at NCBI




## Comparative genomics resources

| | |
|---|---|
| UCSC Genome4 Bioinformatics | genome.ucsc.edu |
| Ensembl | www.ensembl.org |
| MapViewer | www.ncbi.nlm.nih.gov/mapview |
| VISTA Genome Browser | pipeline.lbl.gov |
| Comparative Regulatory Genomics | corg.molgen.mpg.de |
| GALA | www.bx.psu.edu |
| EnsMart | www.ensembl.org/EnsMart |
| PipMaker and MultiPipMaker | www.bx.psu.edu |
| VISTA server | www-gsd.lbl.gov/vista |
| MAVID server | baboon.math.berkeley.edu/mavid |
| zPicture server | zpicture.dcode.org |
| rVISTA server | rvista.dcode.org |
| COGs | www.ncbi.nlm.nih.gov/COG |
| MOSAIC | genome.jouy.inrafr/mosaic |

## "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 |

## 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

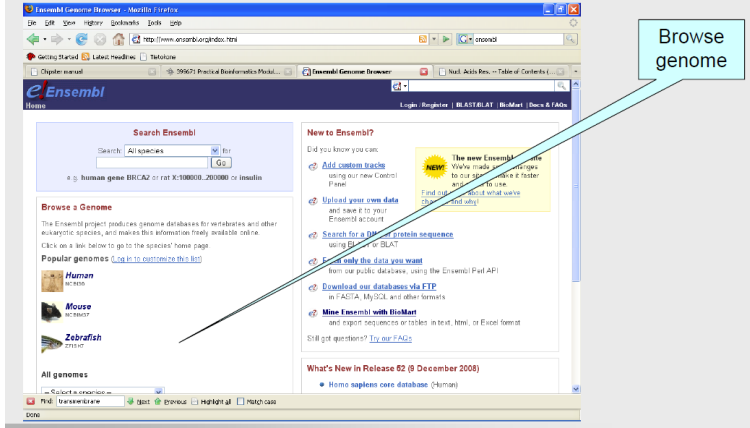
## Why genome databases?

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

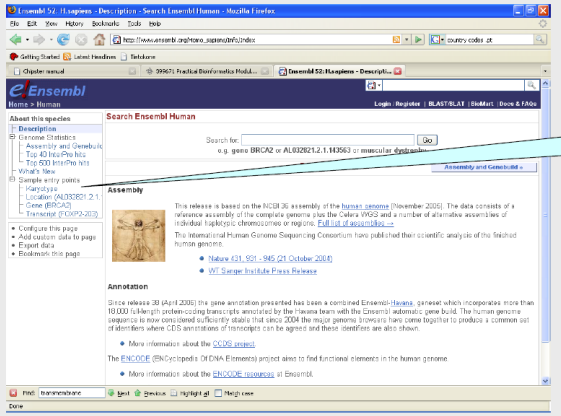
## Some considerations

- **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**

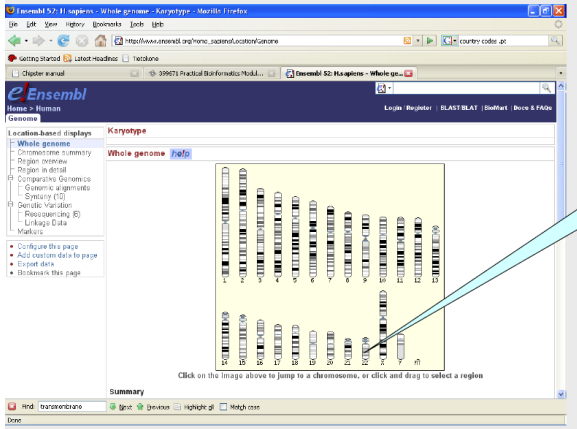
## Ensembl front page



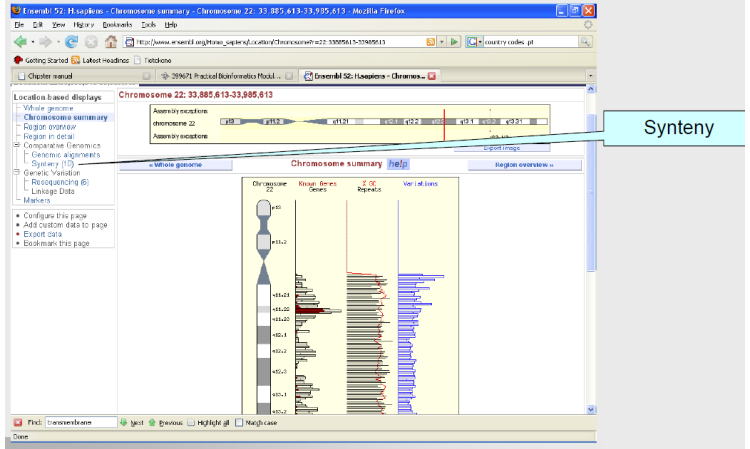

## Explore the genome



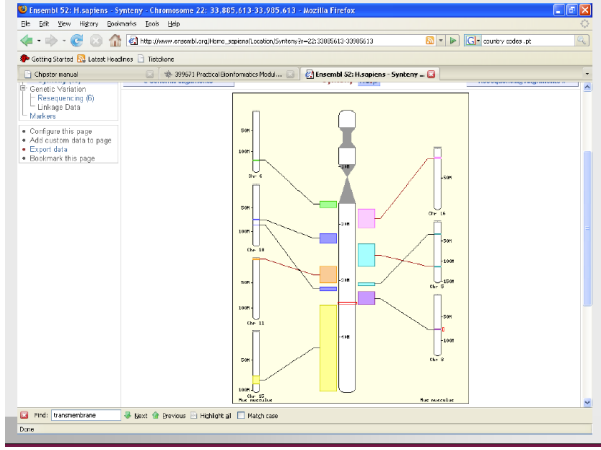

## Explore chromosomes



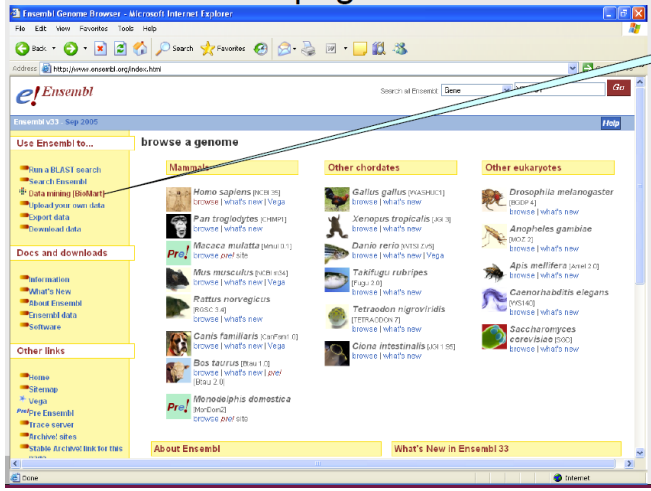

## Chromosome summary



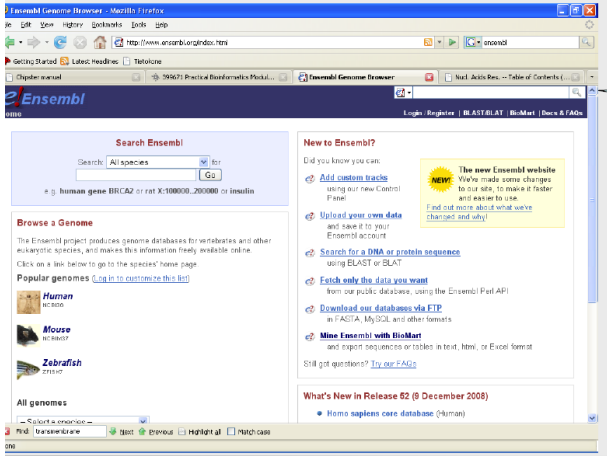

## Synteny View



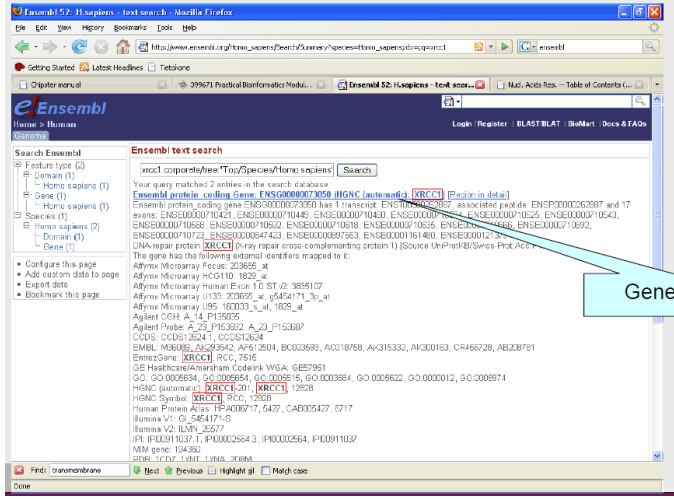
## Ensembl front page



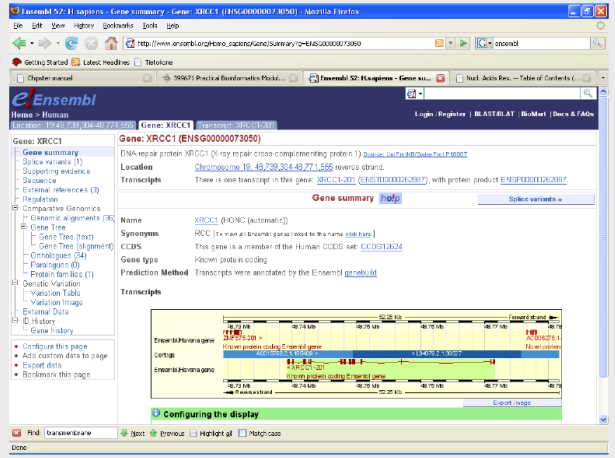
## Ensembl front page



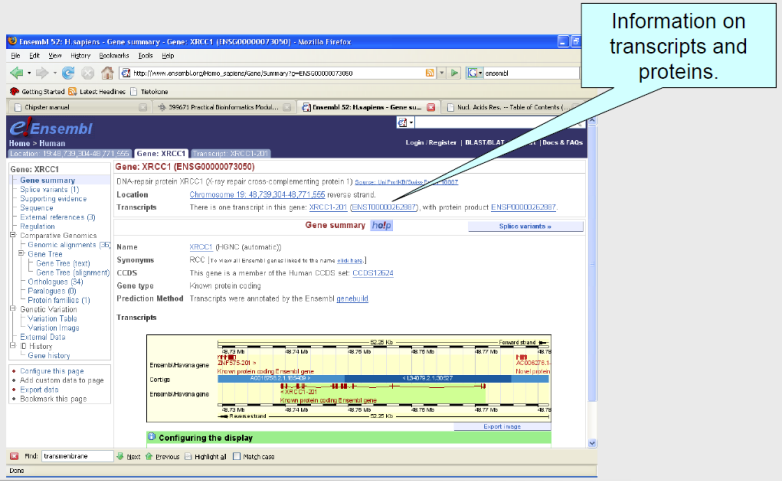

## Quick search results



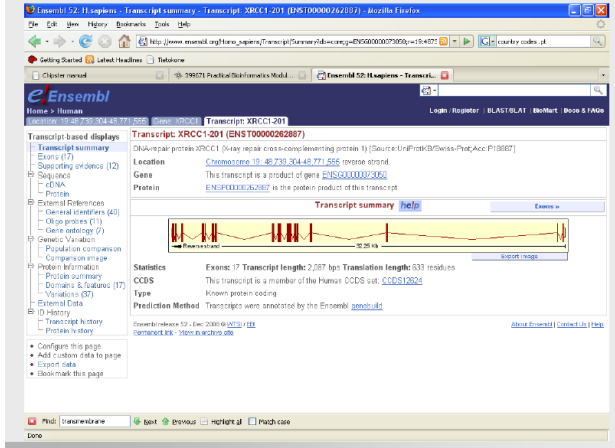

## Gene View



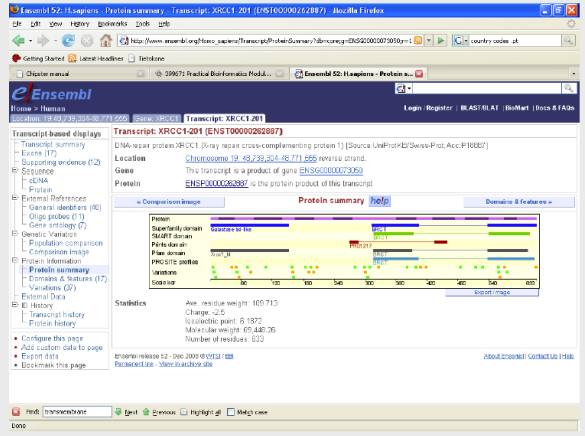
## Gene View



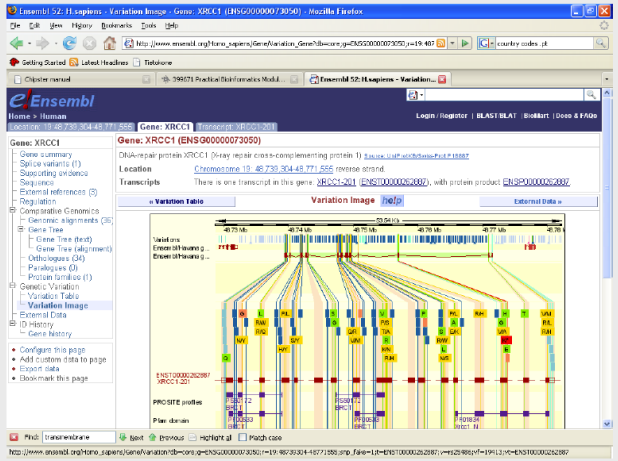

## Transcript info



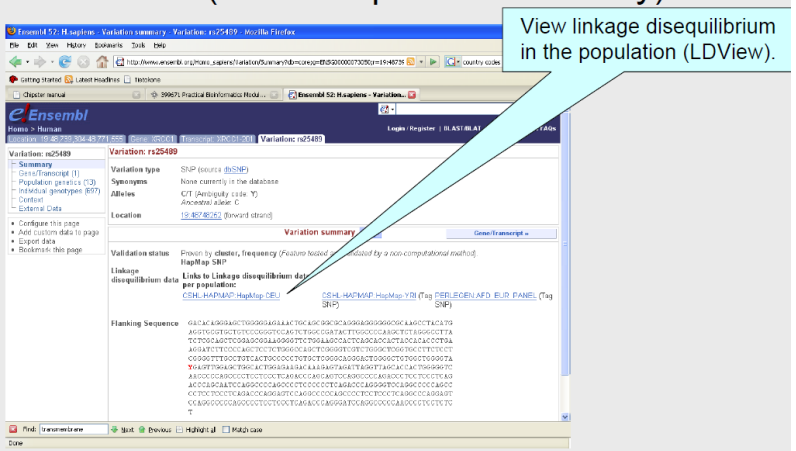
## Protein summary



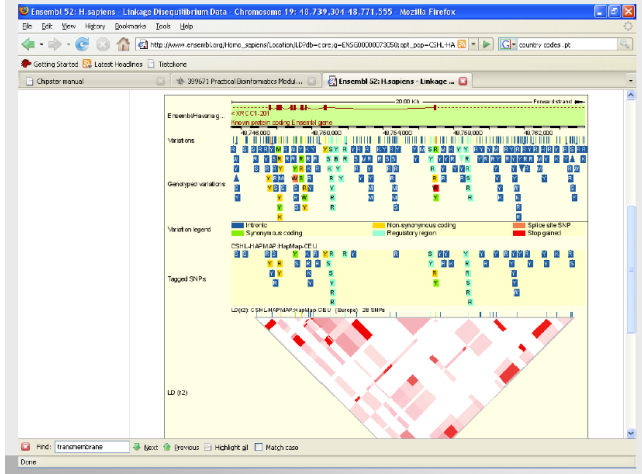
## Variation (SNP) view – from Gene tab



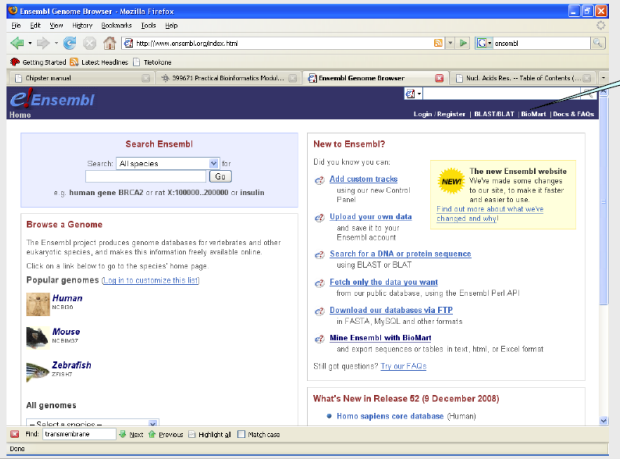
## SNPView (link from protein summary)



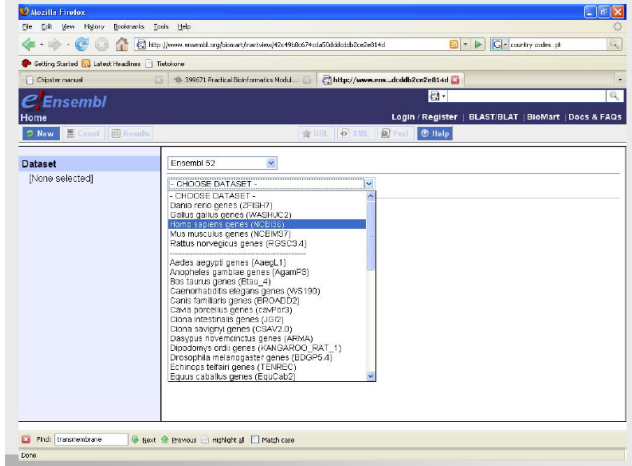

## LDView



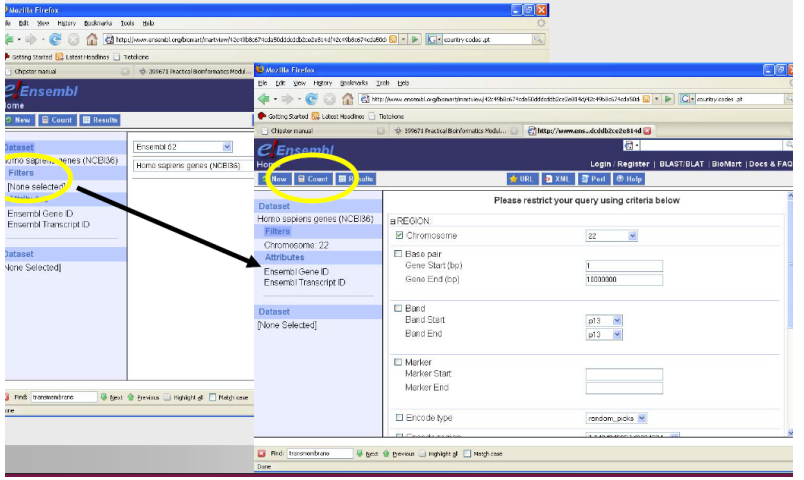
## Ensembl front page



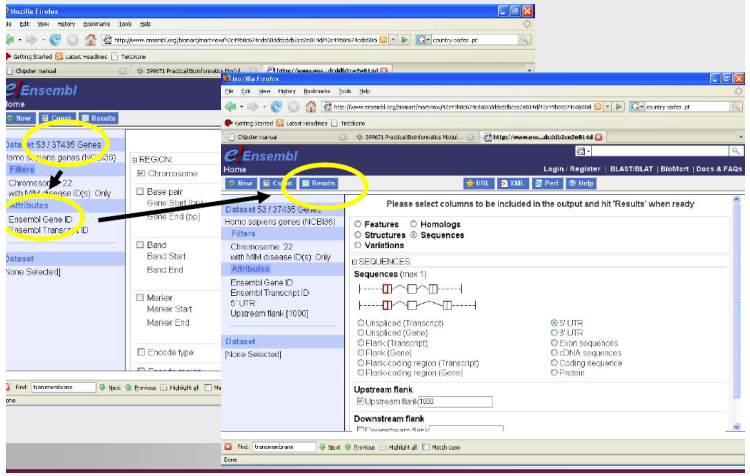

## MartView – select genome



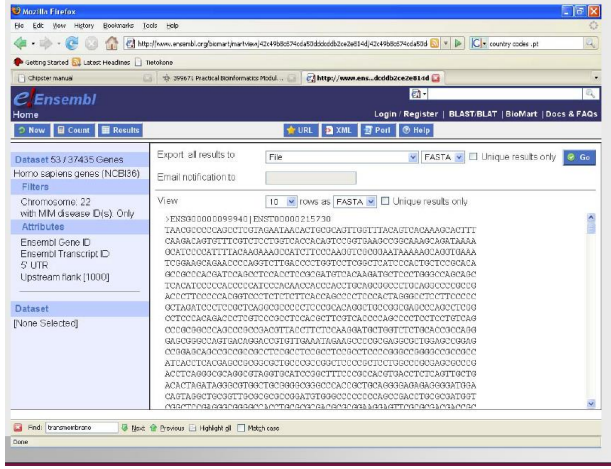
## MartView - Filter



## MartView - output



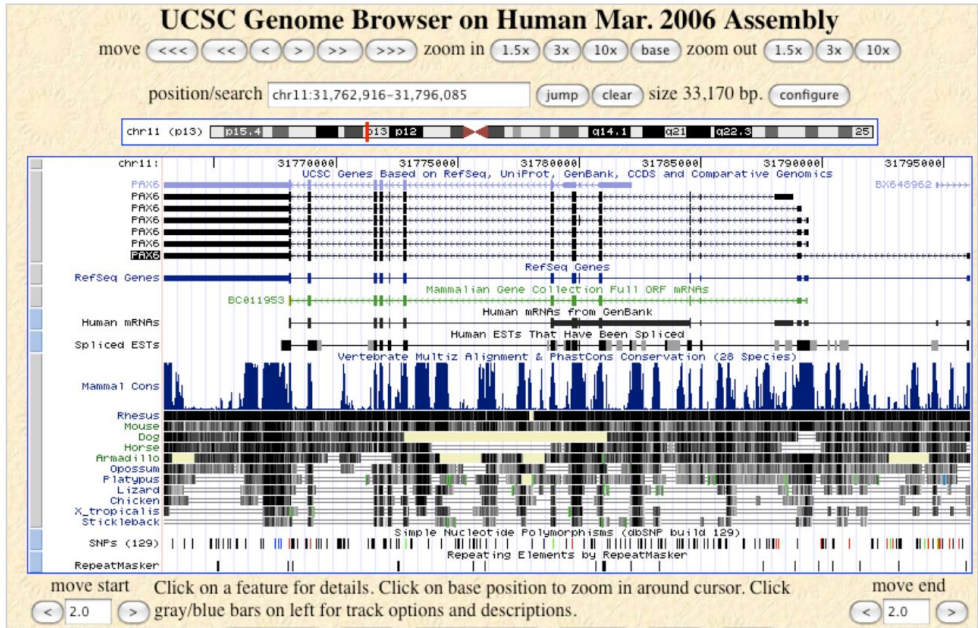
## MartView



## UCSC Genome Browser (University California Santa Cruz - USA)

http://genome.ucsc.edu/

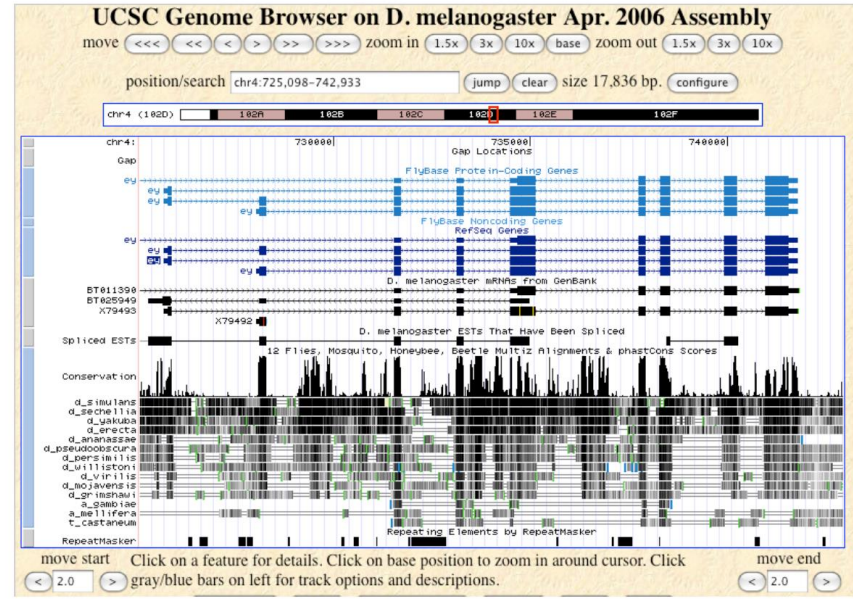
*Human gene Pax6 aligned with Vertebrate genomes*




## UCSC Genome Browser (University California Santa Cruz - USA)

http://genome.ucsc.edu/

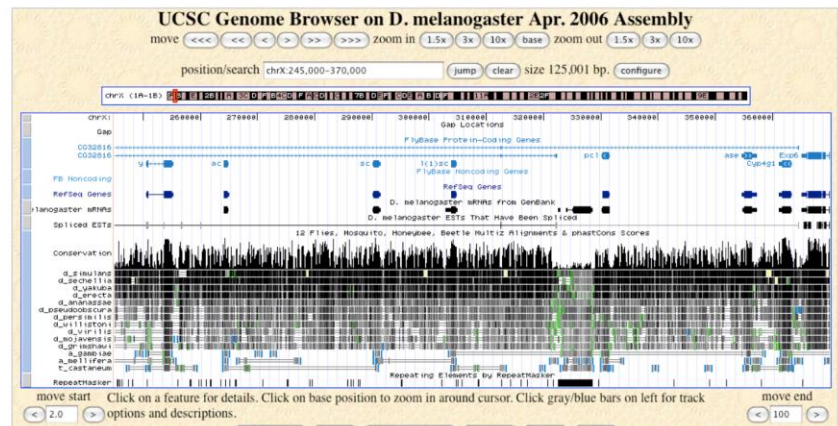
*Drosophila gene eyeless (homolog to Pax6) aligned with Insect genomes*




## UCSC Genome Browser (University California Santa Cruz - USA)

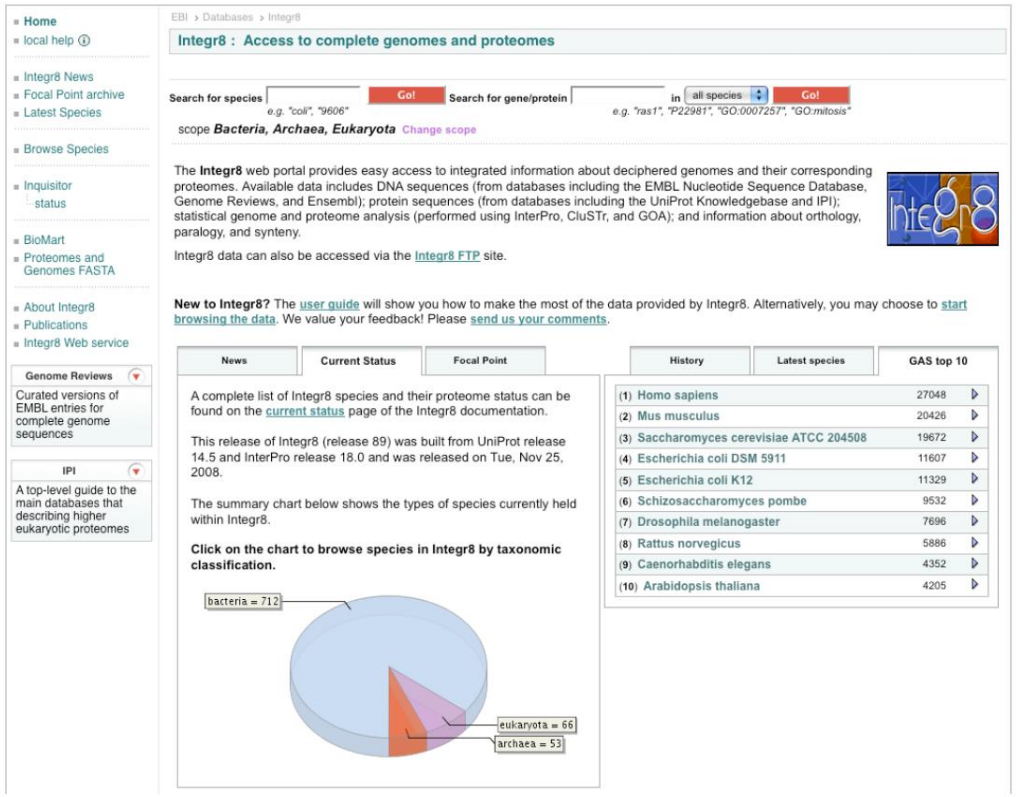
http://genome.ucsc.edu/

*Drosophila 120kb chromosomal region covering the Achaete-Scute Complex*

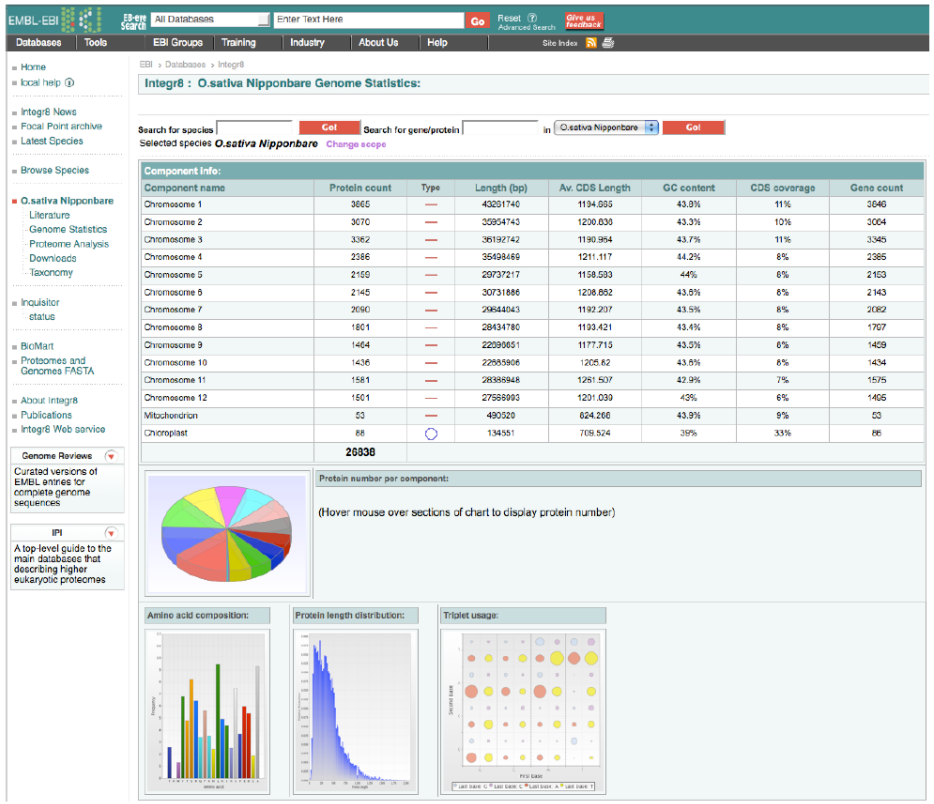


# Comparative genomics

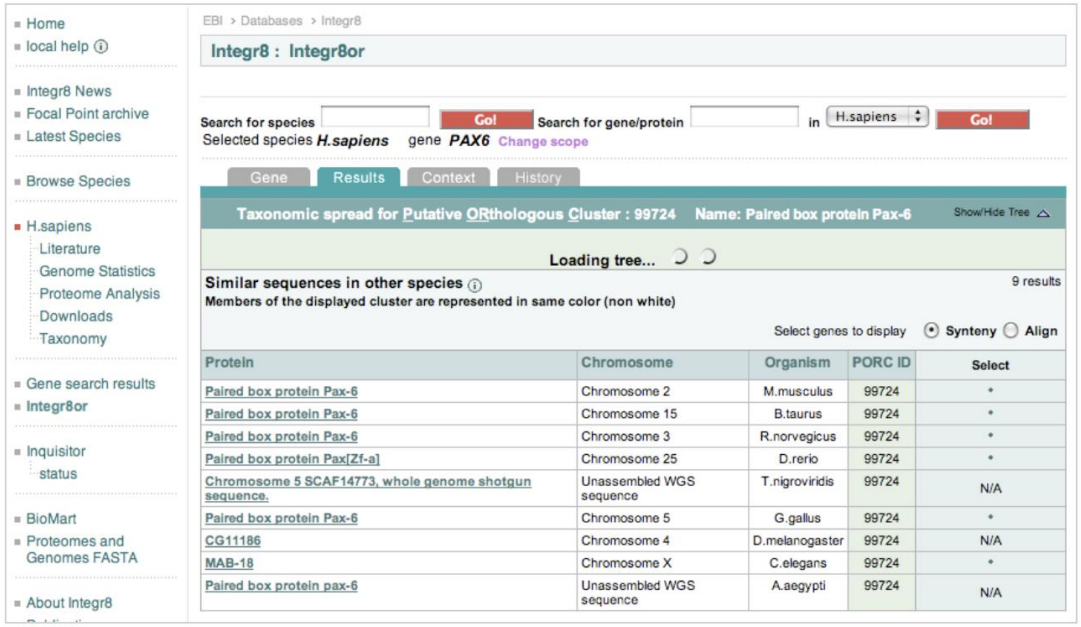
## Integr8 - access to complete genomes and proteomes
*http://www.ebi.ac.uk/integr8/*



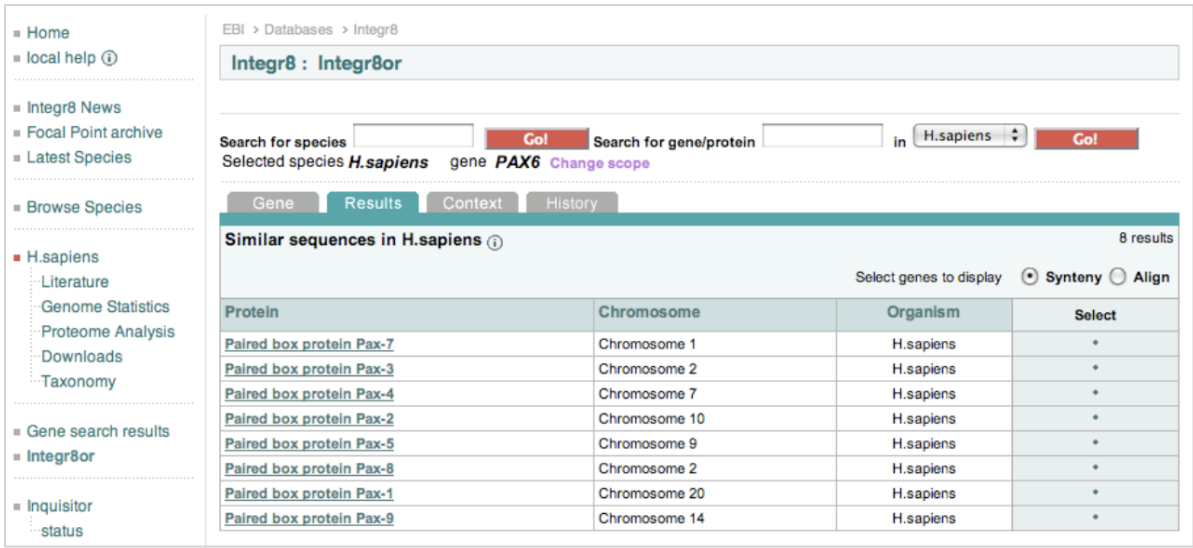

## Integr8 - genome summaries
*http://www.ebi.ac.uk/integr8/*



## *Integr8 - clusters of orthologous genes (COGs)*
*http://www.ebi.ac.uk/integr8/*




## *Integr8 - clusters of paralogous genes*
*http://www.ebi.ac.uk/integr8/*



# 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)<br>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)<br>http://gib.genes.nig.ac.jp/ | Genomes at DDBJ in the Genome Information Broker system |
| European Bioinformatics Institue (EBI)<br>http://www.ebi.ac.uk/genomes/ | Genomes at EBI |
| National Center for Biotechnology Information (NCBI)<br>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)<br>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<br>http://cbi.labri.fr/outils/molligen/ | Website dedicated to mollicute genomes allowing BLAST searching, whole-genome alignment |
| Oral Pathogens database<br>http://www.oralgen.lanl.gov/ | Databases of oral pathogens, bacterial and viral |
| Pathema<br>http://www.tigr.org/pathema/index.shtml | In-depth curatorial analysis of pathogen genomes |
| STDGen and the Oral Pathogens database<br>http://www.stdgen.lanl.gov/ | Databases of genomes responsible for sexually transmitted diseases |
| **Comparative genomic databases** | |
| KEGG: Kyoto Encyclopedia of Genes and Genomes<br>http://www.genome.jp/kegg/ | Enzyme and pathway information about complete genomes |
| Comprehensive Microbial Resource (CMR)<br>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)<br>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)<br>http://mbgd.genome.ad.jp/ | Provides orthologue identification, paralogue clustering, motif analysis and gene order data |
| Virulogenome<br>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)<br>http://www.ncbi.nlm.nih.gov/COG/ | Individual proteins or groups of paralogues from at least three lineages corresponding to ancient conserved domains |
| FusionDB<br>http://igs-server.cnrs-mrs.fr/FusionDB/ | A database of bacterial and archaeal gene fusion events |
| Genome Atlas<br>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)<br>http://www.expasy.org/sprot/hamap/ | HAMAP families are a collection of orthologous microbial protein families, generated manually by expert curators |
| Genome Reviews<br>http://www.ebi.ac.uk/GenomeReviews/ | Up-to-date, standardized and comprehensively annotated view of the genomes |
| Homologous Sequences in Complete Genomes Database<br>http://pbil.univ-lyon1.fr/databases/hogenom.html | Database of homologous genes and access to phylogenetic trees |
| Merops<br>http://merops.sanger.ac.uk/ | Information resource for peptidases and the proteins that inhibit them |

**Table 1.** cont.

| PROKARYOTIC GENOMIC RESOURCES | |
|---|---|
| ORFanage<br>http://www.cs.bgu.ac.il/~nomsiew/ORFans/ | Access to singleton, paralogous and orthologous ORFans in bacterial genomes |
| OrphanMINE<br>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<br>http://www.pathogenomics.bc.ca/IslandPathExamples.html | Identification of horizontally transferred genes and genomics islands, including pathogenicity islands |
| SEED<br>http://theseed.uchicago.edu/FIG/index.cgi | Expert curation of genomic subsystems, or sets of functionally or phenotypically related genes |
| TransportDB<br>http://www.membranetransport.org/ | Database describing the predicted cytoplasmic membrane transport proteins |
| tRNAdb<br>http://lowelab.ucsc.edu/GtRNAdb/ | Genomic tRNA database which contains tRNA identifications made by the program tRNAscan-SE |
| **Pathway and protein interaction databases** | |
| BioCyc<br>http://www.biocyc.org/ | A collection of curated databases each of which describes the genome and metabolic pathways of a single organism |
| MetaCyc<br>http://metacyc.org/ | A database of nonredundant, experimentally elucidated metabolic pathways |
| STRING<br>http://string.embl.de/ | A database of known and predicted protein–protein interactions |
| **Mulitple genome alignment tools** | |
| A Genome Comparison Tool (ACT)<br>http://www.sanger.ac.uk/Software/ACT/ | A DNA sequence comparison viewer (usually BLASTN or tBLASTX) based on the Artemis genome visualization tool |
| Mauve<br>http://gel.ahabs.wisc.edu/mauve/ | Multiple genome alignments in the presence of large-scale evolutionary events |
| Multi-LAGAN<br>http://lagan.stanford.edu/lagan_web/index.shtml | One of several packages in the LAGAN tool set for mulitple alignment of genomes |
| MultiPipMaker<br>http://pipmaker.bx.psu.edu/pipmaker/ | Summarizes similarity between multiple sequences using 'percent identity plots' (Pips) |
| Multiple Genome Aligner (MGA)<br>http://bibiserv.techfak.uni-bielefeld.de/mga/ | Computation of multiple genome alignments of large, closely related DNA sequences |
| **Phylogenomics** | |
| PyPhy<br>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)<br>http://igs-server.cnrs-mrs.fr/phydbac/ | Web interactive tool that displays phylogenomic profiles of bacterial protein sequences |
| **Visualization of multiple genomes** | |
| EnteriX<br>http://globin.cse.psu.edu/enterix/ | Visualization tools for bacterial genome alignments |
| Multiple Genome Navigator (MuGeN)<br>http://www-mig.jouy.inra.fr/bdsi/MuGeN/ | Tool for visual exploration of features of multiple genomes |
| **Genomic metadata** | |
| CMR's Genome Properties<br>http://www.tigr.org/Genome_Properties/ | Numerous attributes whose status can be described by numerical values or controlled vocabulary terms |
| GenomeMine<br>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<br>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)<br>http://www.pasteur.fr/~tekaia/sacso.html | Information on base composition, amino acid composition, ancestral duplication, ancestral conservation and organisms' classification |

# Proteomske podatkovne baze

D380–D386 Nucleic Acids Research, 2009, Vol. 37, Database issue
doi:10.1093/nar/gkn762
Published online 26 November 2008

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

Derek Wilson[1,*], Ralph Pethica[2], Yiduo Zhou[2], Charles Talbot[2], Christine Vogel[3], Martin Madera[2], Cyrus Chothia[1] and Julian Gough[2]

**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 superfamiles, 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.

Published online 9 November 2010
Nucleic Acids Research, 2011, Vol. 39, Database issue D427–D434
doi:10.1093/nar/gkq1130

# SUPERFAMILY 1.75 including a domain-centric gene ontology method

David A. de Lima Morais[1,*], Hai Fang[1], Owen J. L. Rackham[1], Derek Wilson[2], Ralph Pethica[1], Cyrus Chothia[2] 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.**

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 9 2014, pages 1236–1240
doi:10.1093/bioinformatics/btu031

*Sequence analysis*

Advance Access publication January 21, 2014

# InterProScan 5: genome-scale protein function classification

Philip Jones[1,2], David Binns[1], Hsin-Yu Chang[1], Matthew Fraser[1], Weizhong Li[1], Craig McAnulla[1], Hamish McWilliam[1], John Maslen[1,2], Alex Mitchell[1,*], Gift Nuka[1], Sebastien Pesseat[1], Antony F. Quinn[1], Amaia Sangrador-Vegas[1], Maxim Scheremetjew[1], Siew-Yit Yong[1], Rodrigo Lopez[1] 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 reimplementation 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/.

*Published online 18 November 2011*

*Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D71–D75***
*doi:10.1093/nar/gkr981*

# 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

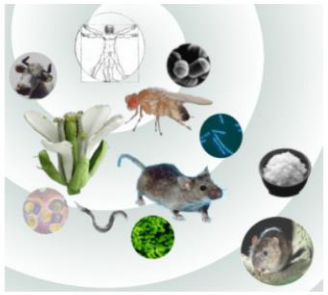
# Bibliographic databases

- **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<br>› Eukaryota<br>› Fungi/Metazoa group<br>› Metazoa<br>› Eumetazoa<br>› Bilateria<br>› Coelomata<br>› Deuterostomia<br>› Chordata<br>› Craniata<br>› 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 ?
  - Assignation of function by similarity
    - recursive process → propagation of errors

## 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

## 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

## 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 an install it in-house (+ add their own private data)
    - academia-industry interface is often ensured by a spinoff company

# 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

# A final rant …

- 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