

Domaine SNV : Biologie, Agronomie, Science Alimentaire, Ecologie

Introduction à la Bioinformatique

www.facebook.com/ DomaineSNV/

Les bases de données biomoléculaires

Jacques van Helden

Jacques.van-Helden@univ-amu.fr

Aix-Marseille Université (AMU), France

Lab. Technological Advances for Genomics and Clinics

(TAGC, INSERM Unit U1090)

<http://jacques.van-helden.perso.luminy.univ-amu.fr/>

Introduction

- Les bases de données jouent un rôle crucial dans l'organisation des connaissances biologiques.
- Nous proposons ici un tour rapide des principales bases de données consultées pour élaborer ce cours, et utilisées durant les travaux pratiques.

Bases de données biomoléculaires

Un tour d'horizon de quelques bases de données biomoléculaires

Exemples de bases de données biomoléculaires

- Séquence et structure des macromolécules
 - Séquences protéiques ([UniProt](#))
 - Séquences nucléotidiques (EMBL / [ENA](#), [Genbank](#), [DDBJ](#))
 - Structures tridimensionnelles des protéines ([PDB](#))
 - Motifs structurels ([CATH](#))
 - Motifs dans les séquences ([PROSITE](#), [PRODOM](#))
- Génomes
 - Bases de données génériques ([Ensembl](#), [UCSC](#), [Integr8](#), [NCBI genome](#), ...)
 - Bases de données spécifiques d'un organisme (SGD, FlyBase, AceDB, PlasmDB, ...)
- Fonctions moléculaires
 - Fonctions enzymatiques, catalyses ([Expasy](#), [LIGAND/KEGG](#), [BRENDA](#))
 - Régulation transcriptionnelle ([JASPAR](#), TRANSFAC, [RegulonDB](#), ...)
- Processus biologiques
 - Voies métaboliques ([MetaCyc](#), [KEGG pathways](#), [Biocatalysis/biodegradation](#))
 - Interactions protéine-protéine ([DIP](#), BIND, [MINT](#))
 - Transduction de signal (Transpath)
 - ...

Bases de données de bases de données

- Il existe des centaines de bases de données spécialisées pour la biologie moléculaire et la biochimie. Ce nombre augmente chaque année.
- Pour s'y retrouver, la revue Nucleic Acids Research consacre chaque année son numéro de janvier à une revue des bases de données existantes, et maintient un catalogue des bases de données:
 - <http://www.oxfordjournals.org/nar/database/c/>
- Plusieurs centaines de bases de données sont disponibles.



Bases de données biomoléculaires

***Bases de données d'acides nucléiques:
GenBank, EMBL, and DDBJ***

Banques de séquences généralistes

- Atlas of Protein Sequences (Margaret Dayhoff)
 - Compilation des séquences à partir de 1965
 - En 1965, contient 50 entrées
 - 1978 : Dernière impression
(ensuite, la base de données sera disponibles sous forme électronique)

Banques généralistes de séquences nucléotidiques

- EMBL (European Molecular Biology Laboratory) :
 - Création 1980 par l' European Molecular Biology Organisation
 - Diffusée par European Bioinformatics Institute (EBI)

- Genbank
 - Création 1982 par IntelliGenetics
 - Diffusée par National Center for Biotechnology Information (NCBI)

- DDBJ (DNA Databank of Japan)
 - Création 1986 par National Institute of Genetics (NIG)
 - Diffusée par National Institute of Genetics (NIG)

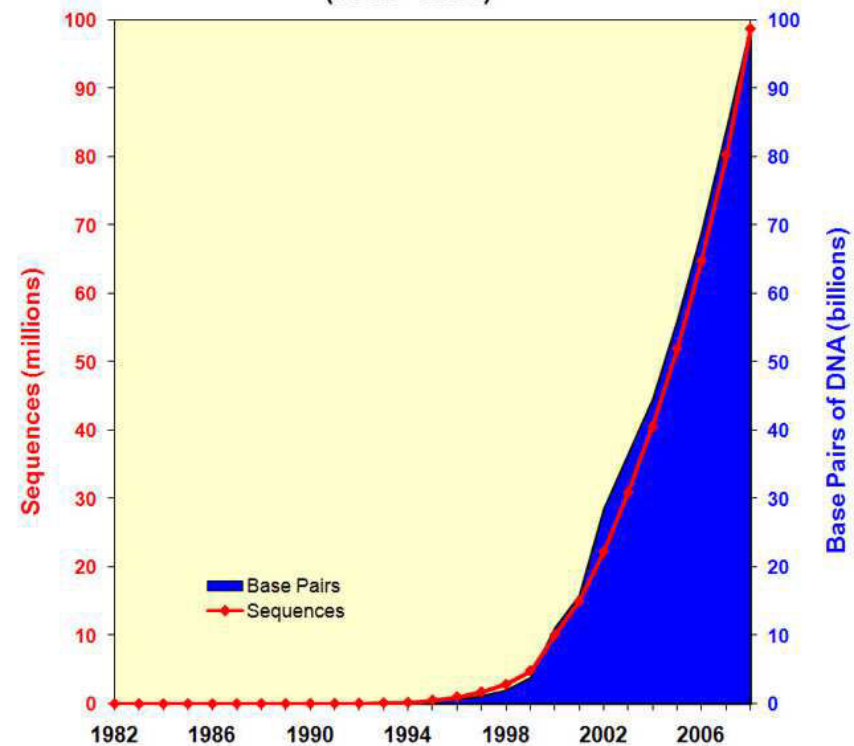
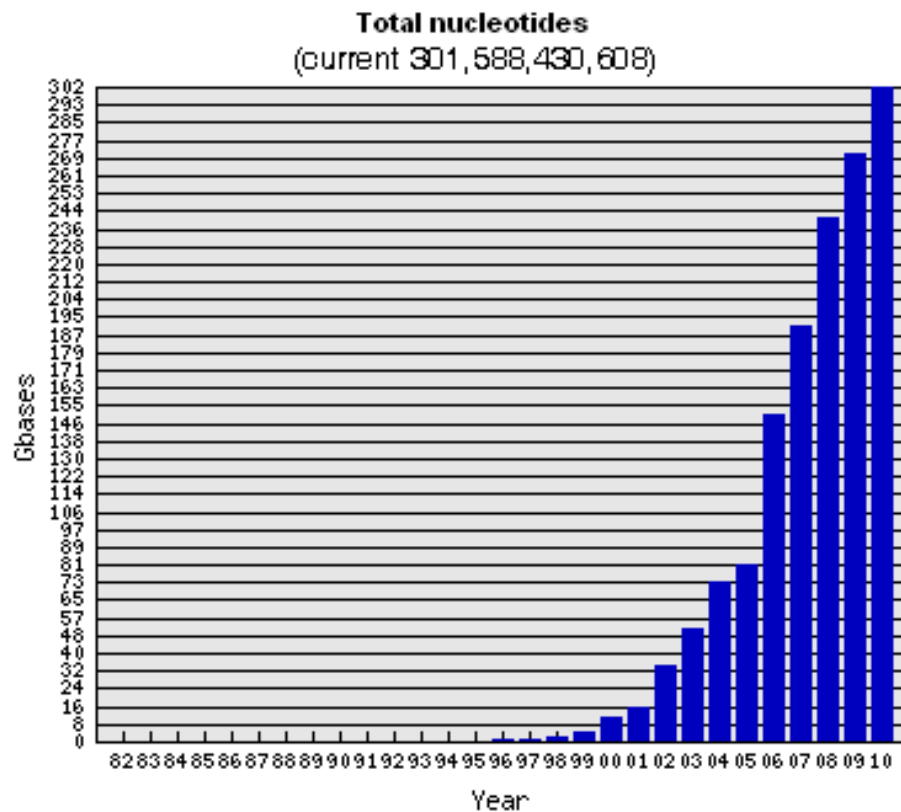
- Ces trois banques échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes
 - « The DDBJ/EMBL/Genbank Feature Table Definition »

Croissance (EMBL, GenBank)

EMBL
300 Gb (décembre 2010)

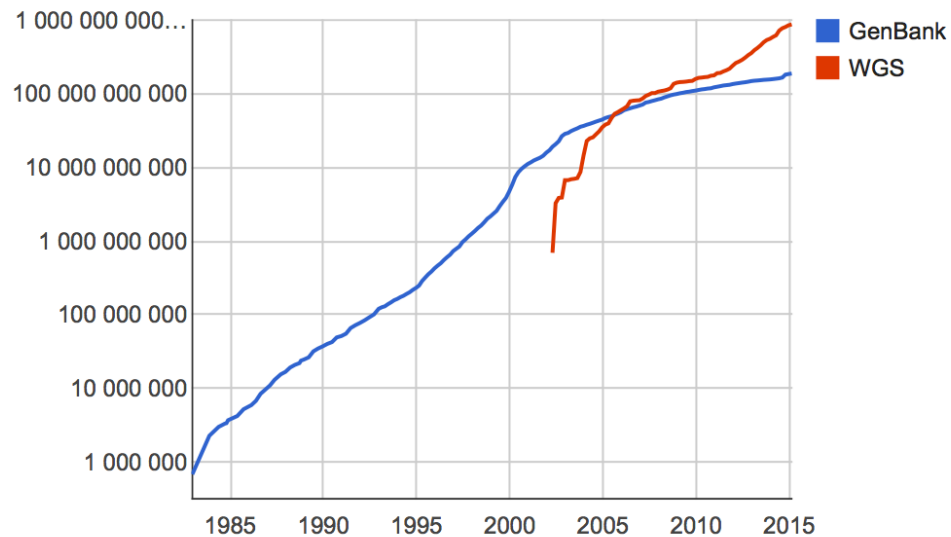
Genbank
286 Gb (août 2010)

<http://www.ebi.ac.uk/embl/Services/DBStats/> <http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html>



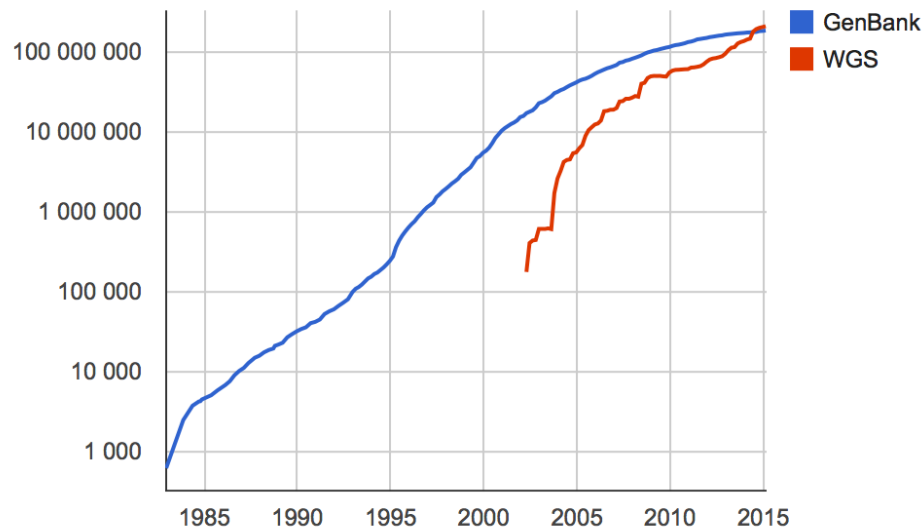
Genbank statistics -2015

Bases

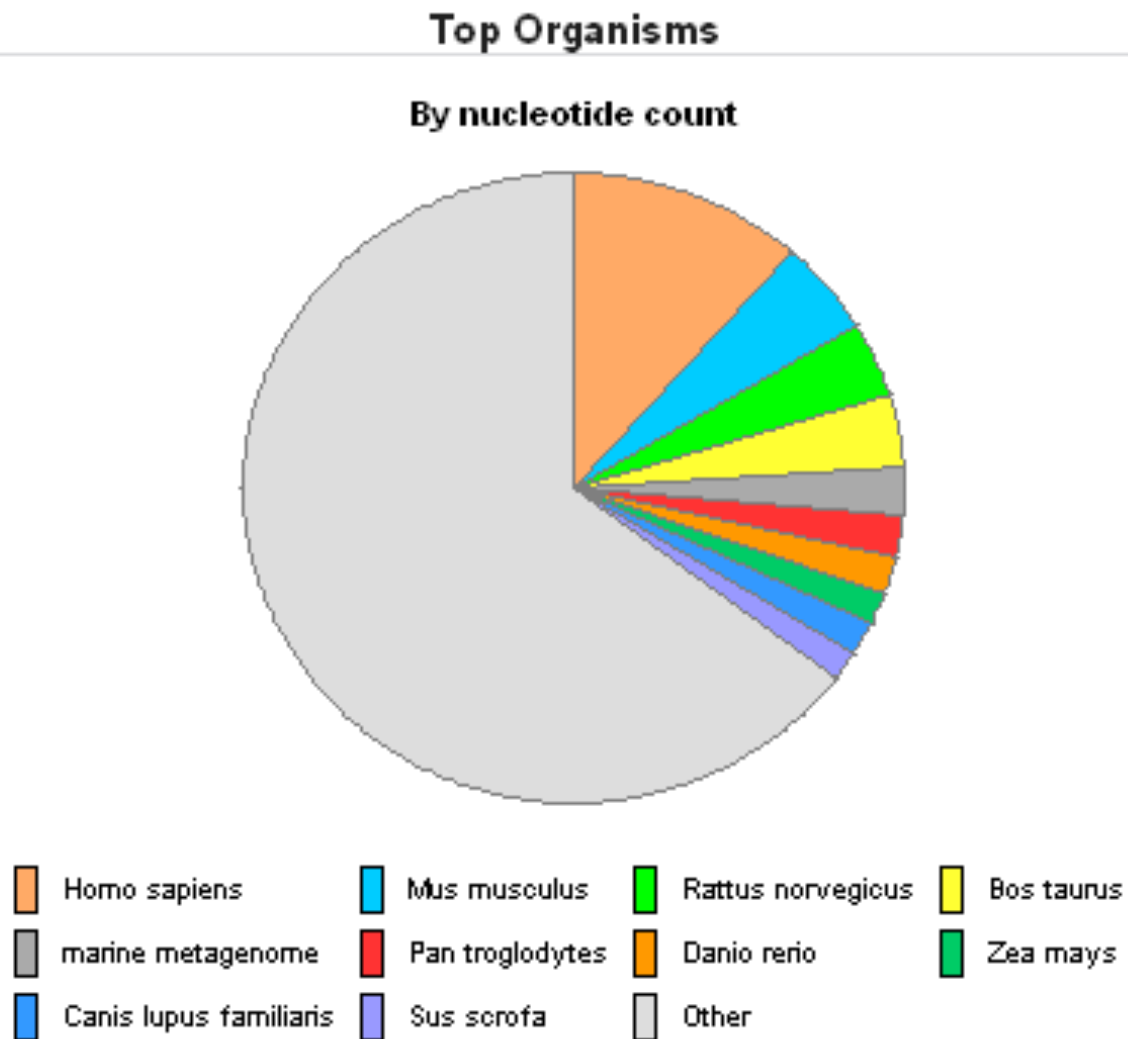


- La taille des données entreposées augmente de façon exponentielle.
- Depuis le début des années 2000, les projets de séquençage génomique constitue une proportion croissante des séquences déposées dans les bases de données.

Sequences

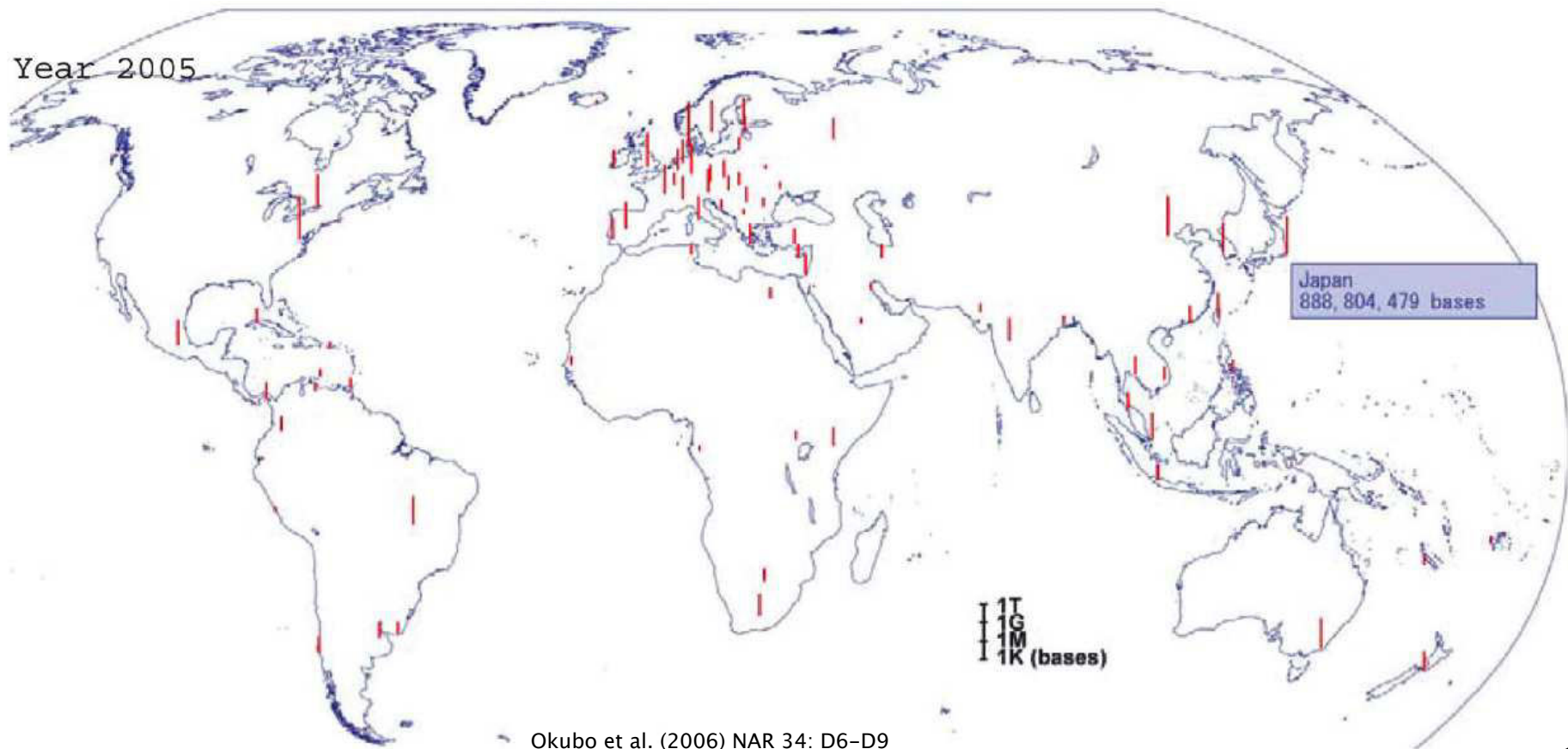


Organismes les plus représentés



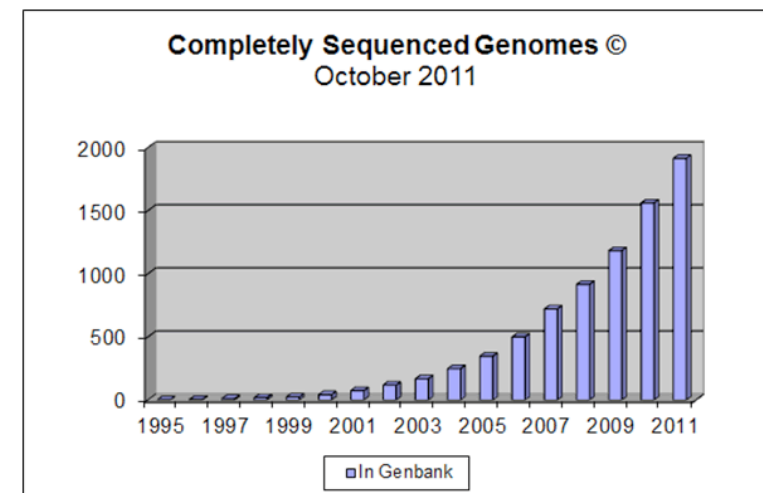
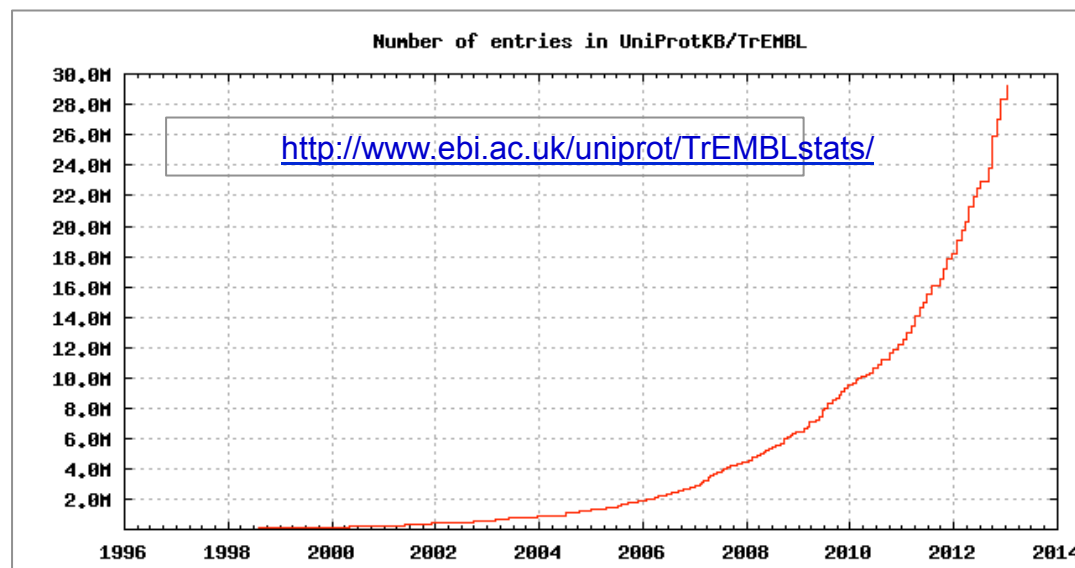
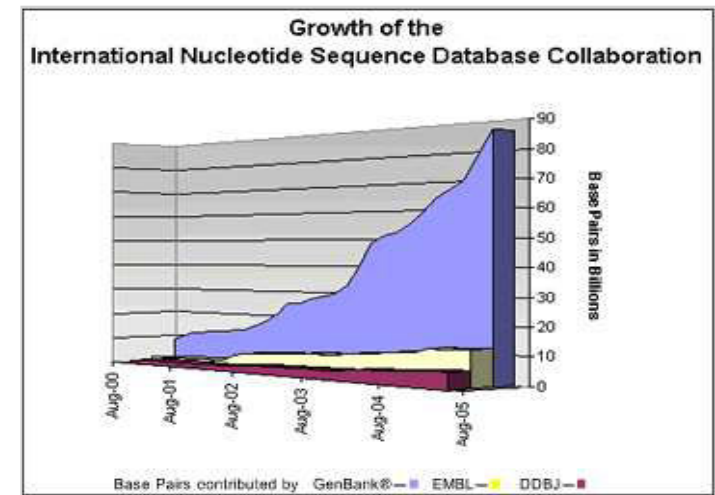
Bases de données de séquences nucléiques

- Avant de publier un article qui décrit une séquence biologique, il est obligatoire de déposer cette séquence dans l'une des 3 principales bases de données de séquences: Genbank (Etats-Unis), EMBL (Europe) ou DDBJ (Japon).
- Les séquences sont automatiquement synchronisées entre les trois bases de données.



L'augmentation exponentielle du nombre de séquences

- Séquences nucléiques
 - Genbank (April 2011) <http://www.ncbi.nlm.nih.gov/genbank/>
 - 135,440,924 séquences totalisant 126,551,501,141 bases dans la division « traditionnelle ».
 - 62,715,288 séquences totalisant 191,401,393,188 bases dans la section « Génomes ».
- Génomes entiers
 - Genomes Online Database (22/01/2013) contient 20531 génomes.
 - <http://www.genomesonline.org/>
- Séquences protéiques
 - Essentiellement toutes proviennent de la traduction (informatique) de séquences nucléiques.
 - UniProtKB/TrEMBL (24/01/2013): 29,266,939 millions de protéines.



Taille des bases de données nucléiques

EMBL Nucleotide Sequence Database: Release Notes - Release 113 September 2012

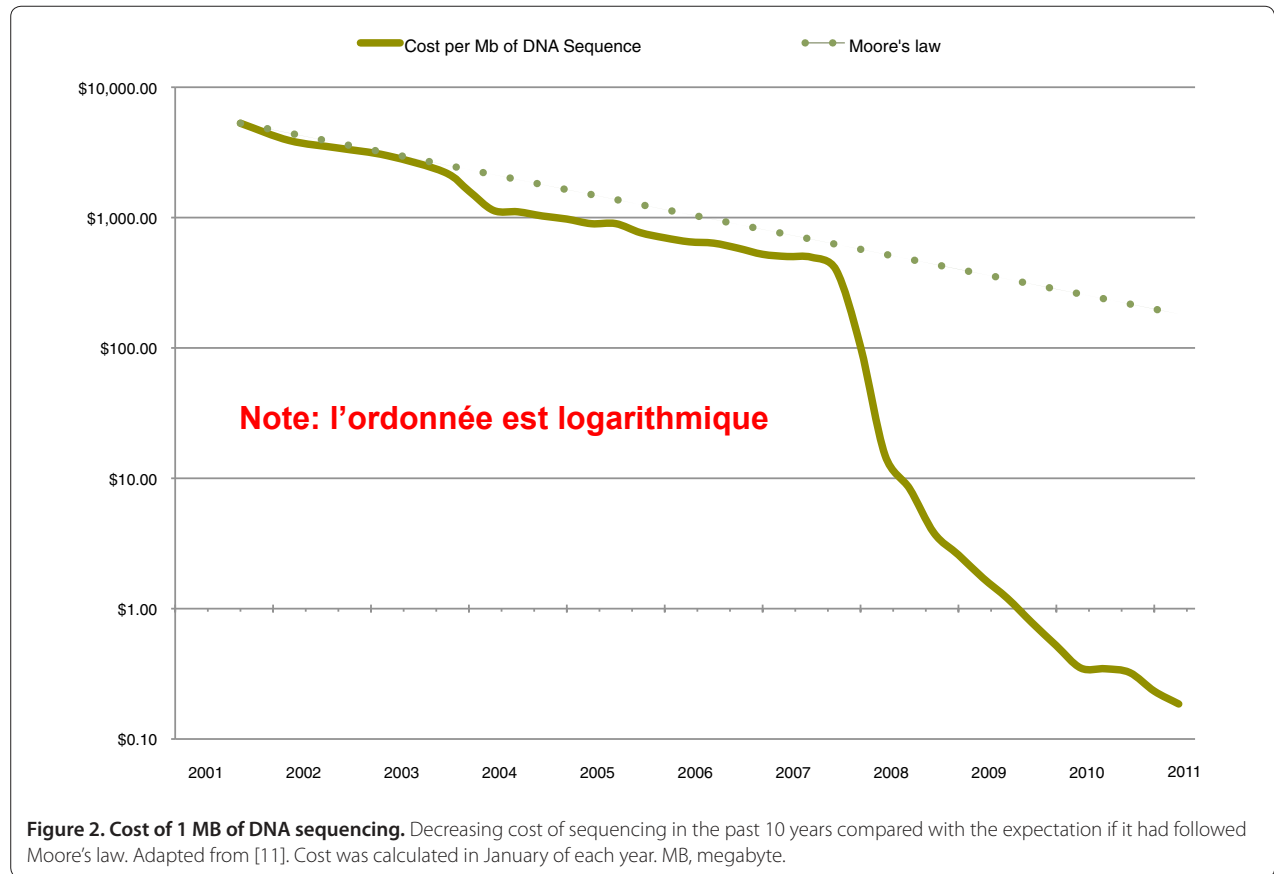
http://www.ebi.ac.uk/embl/Documentation/Release_notes/current/relnotes.html

Class	entries	nucleotides
CON:Constructed	7,236,371	359,112,791,043
EST:Expressed Sequence Tag	73,715,376	40,997,082,803
GSS:Genome Sequence Scan	34,528,104	21,985,922,905
HTC:High Throughput CDNA sequencing	491,770	594,229,662
HTG:High Throughput Genome sequencing	152,599	25,159,746,658
PAT:Patents	24,364,832	12,117,896,594
STD:Standard	13,920,617	37,665,112,606
STS:Sequence Tagged Site	1,322,570	636,037,867
TSA:Transcriptome Shotgun Assembly	8,085,693	5,663,938,279
WGS:Whole Genome Shotgun	88,288,431	305,661,696,545
Total	252,106,363	450,481,663,919

Division	entries	nucleotides
ENV:Environmental Samples	30,908,230	14,420,391,278
FUN:Fungi	6,522,586	11,614,472,226
HUM:Human	32,094,500	38,072,362,804
INV:Invertebrates	31,907,138	52,527,673,643
MAM:Other Mammals	40,012,731	145,678,620,711
MUS:Mus musculus	11,745,671	19,701,637,499
PHG:Bacteriophage	8,511	85,549,111
PLN:Plants	52,428,994	55,570,452,118
PRO:Prokaryotes	2,808,489	28,807,572,238
ROD:Rodents	6,554,012	33,326,106,733
SYN:Synthetic	4,045,013	782,174,055
TGN:Transgenic	285,307	849,743,891
UNC:Unclassified	8,617,225	4,957,442,673
VRL:Viruses	1,358,528	1,518,575,082
VRT:Other Vertebrates	22,809,428	42,568,889,857
Total	252,106,363	450,481,663,919

A l'ère du «Next Generation Sequencing » (NGS)


- Le coût du séquençage décroît de façon exponentielle, et le nombre de séquences augmente de façon exponentielle.
- Jusqu'en 2007, cet effet était compensé par la décroissance exponentielle du coût des ordinateurs (stockage des données, calculs).
- En 2007, plusieurs compagnies ont inventé des nouvelles techniques de séquençage qui ont drastiquement réduit les coûts, et accéléré la production de séquences.
- Le coût du séquençage décroît de façon beaucoup plus rapide que celui du stockage, ce qui commence à créer de vrais problèmes de gestion des données.



Sboner et al. (2011) The real cost of sequencing: higher than you think!. Genome Biol 12: 125

Genbank (NCBI - USA)

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

**NCBI**

National Center for Biotechnology Information
[National Library of Medicine](#) [National Institutes of Health](#)

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search for

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

DatabasesToolsEBI GroupsTrainingIndustryAbout UsHelpSite 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

ation

submissions ...[more](#)

Collaborations

- [INSDC](#) - International Nucleotide Sequence Database Collaboration
- [NCBI](#) - The Nucleotide Sequence Database is produced in

EBI > Databases > EMBL-Bank

EMBL Nucleotide Sequence Database

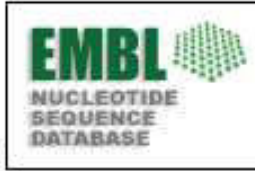
The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are [direct submissions](#) from individual researchers, genome sequencing projects and patent applications.


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 , DE 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

Accession [DNA](#) [Protein](#) [Taxonomy](#) [Site Search](#)

Accession numbers

☒ DDBJ
 ☐ UniProt
 ☐ PDB
 ☐ DAD
 ☐ 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 [▶ More](#)

- ▶ 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 [▶ More](#)


- ▶ 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 \(\[ftp.ddbj.nig.ac.jp\]\(ftp://ftp.ddbj.nig.ac.jp\) \)](#)

Download data files

Bases de données biomoléculaires

Bases de données de séquences protéiques

Banques généralistes de séquences protéiques

- PIR (Protein Information Resource)
 - Première banque des protéines (1965)
 - Banque américaine (NBRF- National Biomedical Research Foundation)
 - Protéines regroupés en familles
 - Sur base de similarité de séquences, on peut identifier des familles de protéines, et extraire ainsi les données une connaissance plus spécifique que la somme des informations contenues dans les séquences individuelles.

Banques généralistes de séquences protéiques

- SwissProt

- Projet démarré par Aimos Bairoch en 1986, à l'université de Genève
- Chaque séquence est expertisée par un « annotateur » (ou « curateur »), expert dans un domaine particulier de la biologie.
- Réellement non-redondante: quand plusieurs séquences sont identiques, une séquence est choisie comme « représentative ». Les séquences proches sont recensées en tant que variants

Banques généralistes de séquences protéiques

- TrEMBL : traduction automatique de EMBL
- Genpept traduction automatique de GenBank
- Etant donné l'augmentation exponentielle des séquences nucléiques disponibles, ces bases de données de protéines traduites augmentent aussi de façon exponentielle.
 - En Novembre 2014, TrEMBL contient >86 millions de séquences
 - Swissprot (annotations) n'en contient que 546.790.
- Il devient impossible de réviser une par une ces séquences. Elles sont donc annotées automatiquement:
 - Identification des domaines sur base de similarités de séquences.
 - Annotation de la fonction de la protéine sur base de similarité de séquence.
 - Ces annotations sont bien entendu sujettes à caution : l'annotation automatique présente des risques d'erreurs
 - échec d'identification d'un domaine ou d'une fonction,
 - assignation erronée d'un domaine ou d'une fonction.

Swiss-Prot + PIR + TrEMBL-EBI



UniProt

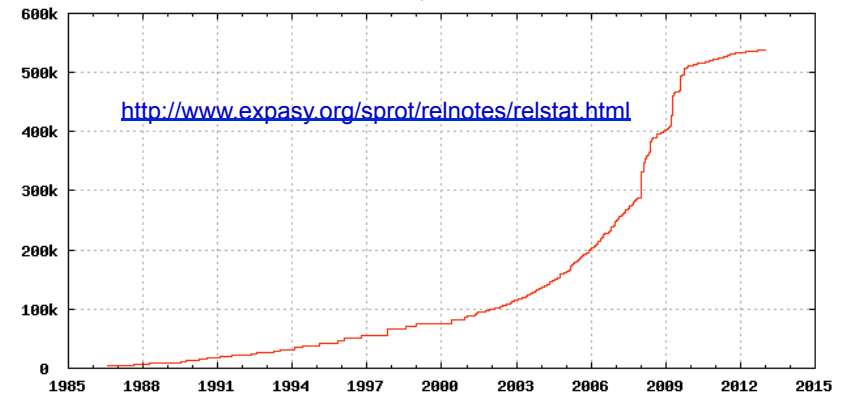
(Universal Protein Ressource)

<http://www.uniprot.org/>

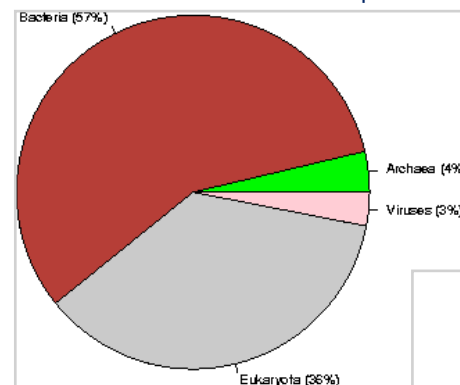


- Contenu (Novembre 2014)
 - UniProtKB (Swissprot + TrEMBL)
 - **86.536.393 protéines**
 - Traduction et annotation automatique de toutes les séquences codantes d'EMBL
 - Section Swiss-Prot d'UniProtKB (« reviewed »):
 - **546.790 protéines**
 - annotation par des experts
 - Contenu informationnel important.
 - Nombreuses références à la littérature scientifique.
 - Bonne fiabilité des informations.
 - La majorité des annotations de séquences protéiques sont donc faites automatiquement, sans être vérifiées par un être humain !!!
- Swissprot
 - La base de données de protéines la plus complète au monde.
 - Une énorme équipe: >100 annotateurs + développeurs d'outils.
 - Annotation par experts, spécialistes des différents types de protéines.
- 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.

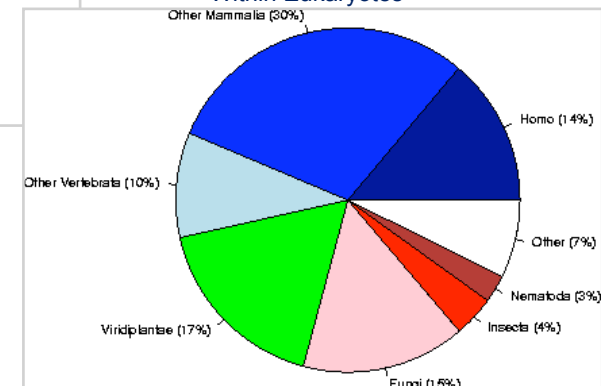
Number of entries (polypeptides) in Swiss-Prot



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	<i>Recommended name:</i> Paired box protein Pax-6 <i>Alternative name(s):</i> Oculorhombin Aniridia type II protein
Gene names	Name: PAX6 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.

7

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

len

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.

Detailed description of regions, variations, and secondary structure

Experimental info					
<input type="checkbox"/>	Sequence conflict	317	1	R → L in AAA59963 and AAA59962 . Ref. 1	
<input type="checkbox"/>	Sequence conflict	369	1	Y → C in CAE45868 . Ref. 4	

UniProt example - Human Pax-6 protein

Peptidic sequence

Sequences						Hide Top
Sequence	Length	Mass (Da)	Tools			
<input type="checkbox"/> Isoform 1 [UniParc]. Last modified July 15, 1999. Version 2. Checksum: C33CDD2C1B13C397	FASTA	422	46,683	Blast	go	
<pre> 10 20 30 40 50 60 MQNSHSGVNQ LGGVFVNGRP LPDSTRQKIV ELAHSGARPC DISRILQVSN GCVSKILGRY 70 80 90 100 110 120 YETGSIRPRA IGGSKPRVAT PEVVSIAQY KRECPSIPAW EIRDRLLESEG VCTNDNIPSV 130 140 150 160 170 180 SSINRVLRLN ASEKQMGAD GMYDKLRMLN GQTGSWGTRP GWYPGTSVPG QPTQDGCQQQ 190 200 210 220 230 240 EGGGENTNSI SSNGEDSDEA QMRLQLKRKL QRNRTSFTQE QIEALEKEFE RTHYPDVFFAR 250 260 270 280 290 300 ERLAAKIDLP EARIQVWPSN RRAKWRREEK LRNQRQASN TPSHIPISSE FSTSVYQPIP 310 320 330 340 350 360 QPTTPVSSFT SGSMGLRTDT ALTNTYSALP PMPSTTMANN LPMQPPVPSQ TSSYSCLMPT 370 380 390 400 410 420 SPSVNGRSYD TYTPPHMQTH MNSQPMGTSG TTSTGLISPG VSVPVQVPQS EPDMSQYWPR LQ </pre>						
« Hide						
<input type="checkbox"/> Isoform 5a (Pax6-5a) [UniParc]. Checksum: 74926827347A20B5 Show »		436	48,218	Blast	go	
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).	
Cited for: VARIANT ANEOTROPY PRO-362.		
[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.
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

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

ENSG000000007372. Homo sapiens. [Contig view]

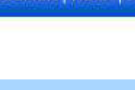
GeneID

5080.

KEGG

hsa:5080.

Structure tridimensionnelle des protéines



PROTEIN DATA BANK

[CONTACT US](#) | [FEEDBACK](#) | [HELP](#) | [PRINT](#)

[PDB ID or keyword](#)
[Author](#)
[Site Search](#)
[Advanced Search](#)

[Home](#)
[Search](#)

- Home
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- General Education
- Site Tutorials
- BioSync
- General Information
- Acknowledgements
- Frequently Asked Questions

Quick Tips:

Want to search by sequence? Click [here](#).

Are you missing data updates? The PDB archive has moved to <ftp://ftp.wwpdb.org>. For more information click [here](#).

Welcome to the RCSB PDB

The **RCSB** PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the **wwPDB** whose mission is to ensure that the PDB archive remains an international resource with uniform data.

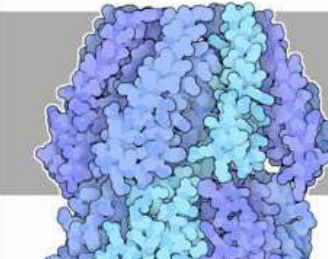
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

A **list of browsers** known to work with this site and a browser compatibility **check** are available.

A **narrated tutorial** illustrates how to search, navigate, browse, generate reports and visualize structures using this site. [Requires the Macromedia [Flash player](#).]

[Comments?](#)

Molecule of the Month: Mechanosensitive Channels



We are remarkably resistant to changes in our surrounding environment. Our bulky bodies allow us to weather extremes of heat and cold, and our skin protects us if we go for a swim in fresh water or salty water. If things get too uncomfortable, we can always get up and walk away, finding a warmer or cooler or drier place. Bacteria don't have as many options. They are tiny and they are immersed in water, so changes in the environment can pose life-threatening challenges. For instance, if it rains they may be suddenly surrounded by fresh water. This is dangerous because the water seeps into the cell through osmosis and increases the pressure inside. At other times, the bacterium may be shifted suddenly to salty conditions, which pulls water out and dehydrates the cell. Bacteria have methods for resisting these changes, so they can keep a steady, comfortable osmotic pressure inside.

[More ...](#)
[Previous Features](#)

News

- [Complete News](#)
- [Newsletter](#)
- [Discussion Forum](#)
- [Job Listings](#)

25-November-2008

Announcement: New Releases to Follow Format Guide Version 3.20

Beginning December 2, 2008, all newly-released PDB entries will follow PDB File Format Contents Guide Version 3.20 ([PDF](#) | [HTML](#)).

[Full article ...](#)


16-September-2008

Announcement: Comprehensive Format Guide Version 3.2

During the past year, the wwPDB annotators have collaborated on a project to clarify the details and procedures related to data processing and annotation.

[Full article ...](#)

***Outils d'exploration des génomes
(« genome browsers »)***


Home

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

Search Ensembl


Search: for


e.g. human gene BRCA2 or rat X:100000..200000 or insulin


Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. Click on a link below to go to the species' home page.

Popular genomes ([Log in to customize this list](#))

 **Human**
NCBI36



 **Mouse**
NCBIM37

 **Zebrafish**
ZFISH7

All genomes

[View full list of all Ensembl species](#)

Other pre-build species are available in [Ensembl Pre!](#)



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.

Ensembl receives major funding from the Wellcome Trust. Our [acknowledgements page](#) includes a list of additional current and previous funding bodies.

New to Ensembl?

Did you know you can:

- [Add custom tracks](#)
using our new Control Panel
- [Upload your own data](#)
and save it to your Ensembl account
- [Search for a DNA or protein sequence](#)
using BLAST or BLAT
- [Fetch only the data you want](#)
from our public database, using the Ensembl Perl API
- [Download our databases via FTP](#)
in FASTA, MySQL and other formats
- [Mine Ensembl with BioMart](#)
and export sequences or tables in text, html, or Excel format

Still got questions? [Try our FAQs](#)

What's New in Release 51 (18 November 2008)

- **Webcode version 4.0 released** (all species)
- **New 2x genomes** (multiple species)
- **New Guineapig assembly and genebuild** (Guinea Pig)
- **Human patch for 51** (Human)
- **New xref sources** (Human)

[More news...](#)

Latest Blog Entries

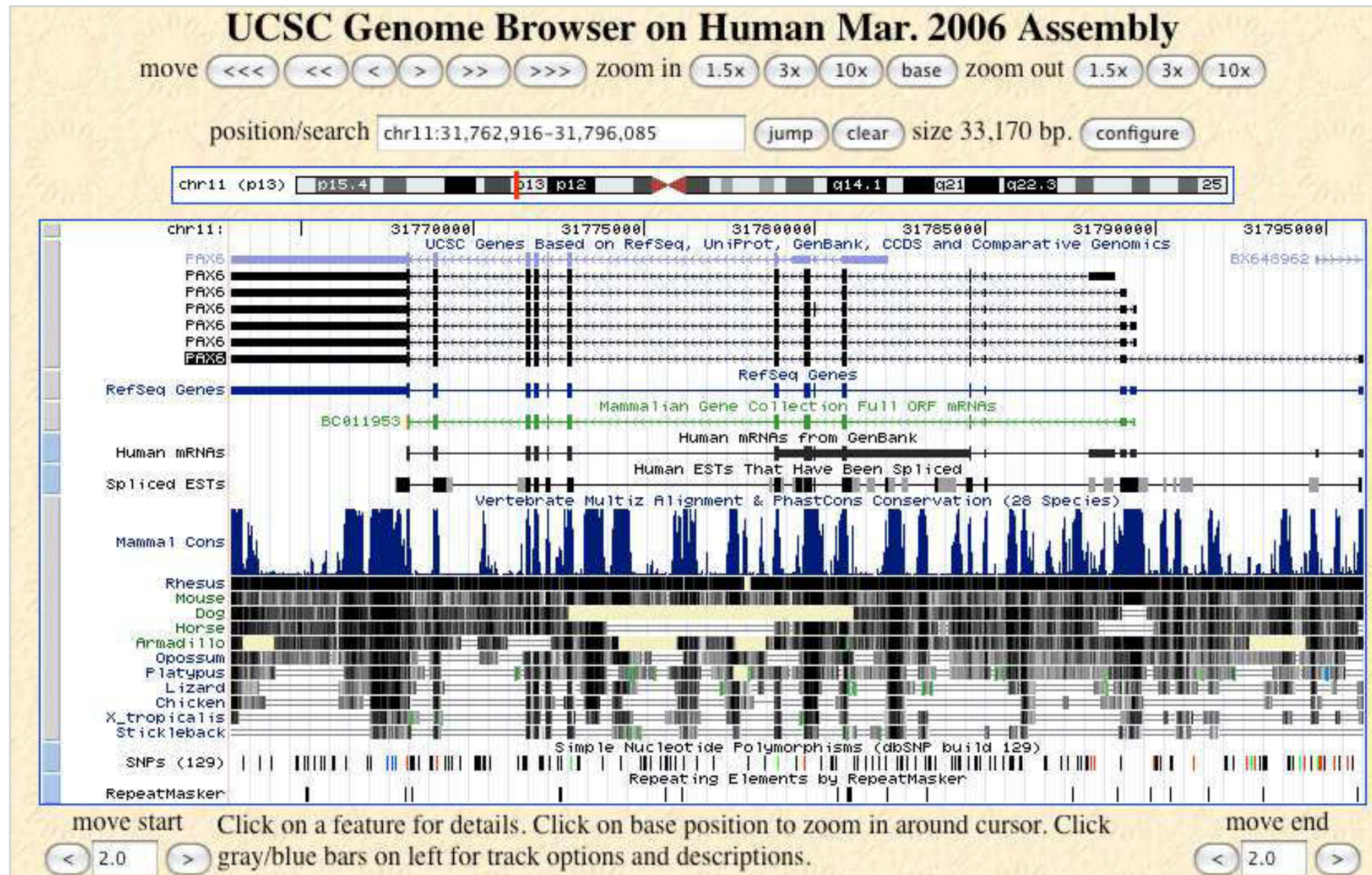
- Sat, 29 Nov 2008: [Ensembl 51](#)
- Wed, 19 Nov 2008: [Upcoming training events December](#)
- Mon, 17 Nov 2008: [Accessing the Ensembl data with Perl](#)

[Go to Ensembl blog](#)

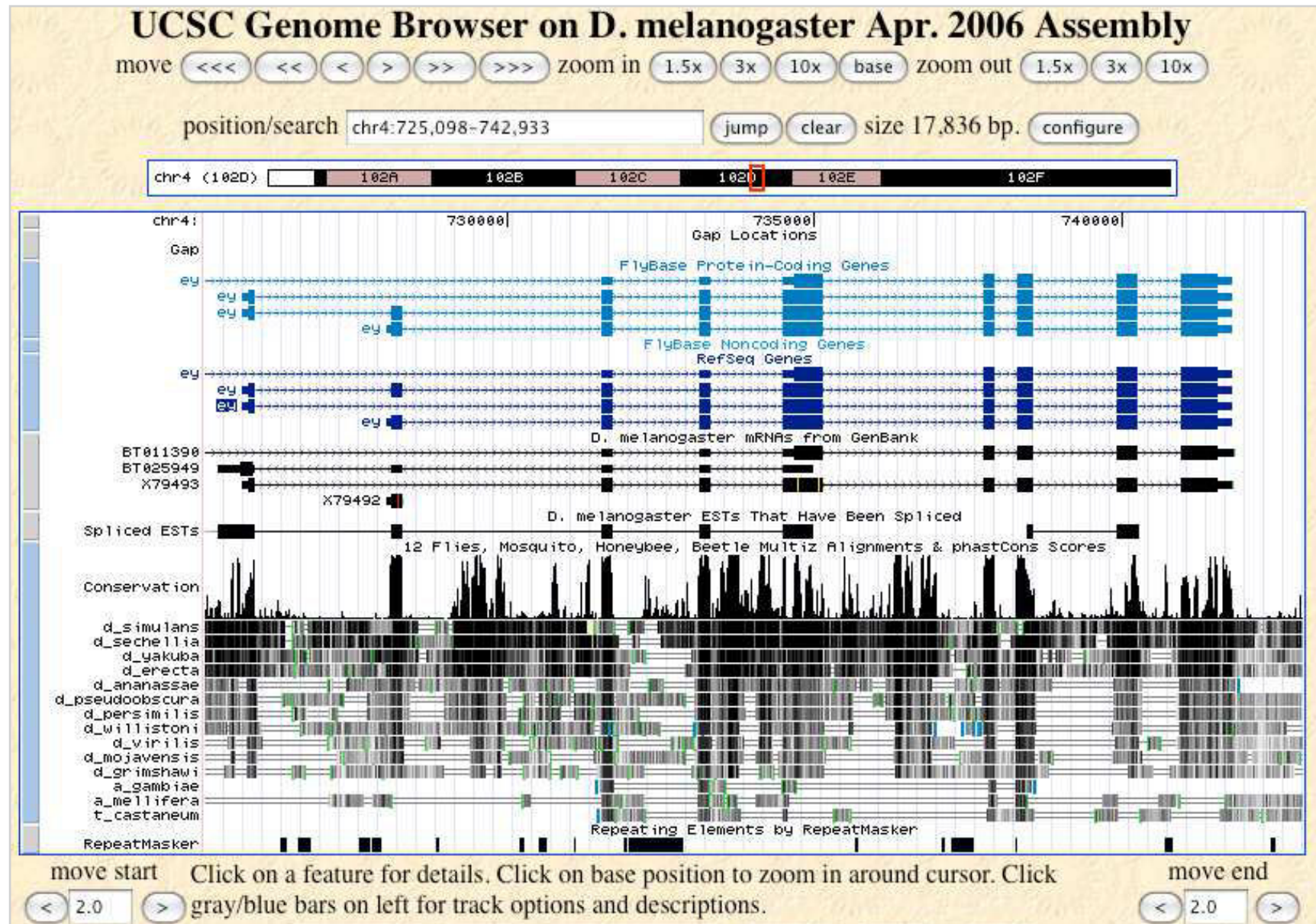
Ensembl release 51 - Nov 2008 © WTSI / EBI
[Permanent link](#) - [View in archive site](#)

[About Ensembl](#) | [Contact Us](#) | [Help](#)

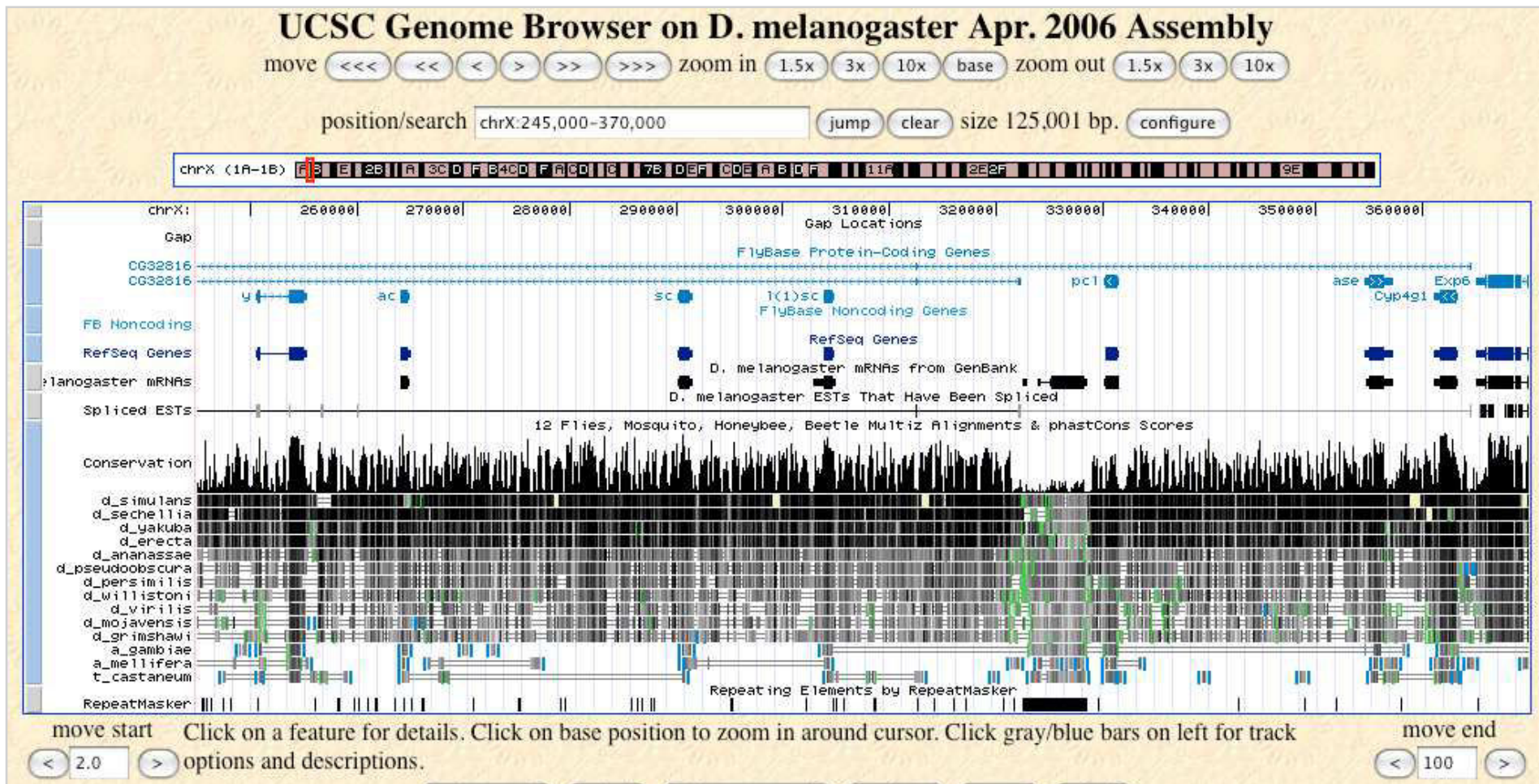
Human gene Pax6 aligned with Vertebrate genomes



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



Drosophila 120kb chromosomal region covering the Achaete-Scute Complex



<http://ecrbrowser.dcode.org/>



EnsEMBL - Example: Drosophila gene Pax6

<http://www.ensembl.org/>



Home > Fruitfly

Location: 4:718,315-741,787 Gene: ey

Gene: ey

- Gene summary
- Splice variants (4)
- Supporting evidence
- Sequence
- External references (55)
- Regulation
- Comparative Genomics
 - Genomic alignments (0)
 - Gene Tree
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (0)
 - Paralogues (0)
 - Protein families (1)
- Genetic Variation
 - Variation Table
 - Variation Image
- External Data
- ID History
 - Gene history
- Export gene data

- Bookmark this page
- Configure this page
- Add custom data to page

Gene: ey (FBgn0005558)

Location [Chromosome 4: 718,315-741,787 forward strand.](#)

Transcripts There are 4 transcripts in this gene: [hide transcripts](#)

ey-RA	FBtr0089236	FBpp0088300	protein_coding
ey-RB	FBtr0089235	FBpp0088299	protein_coding
ey-RC	FBtr0100395	FBpp0099809	protein_coding
ey-RD	FBtr0100396	FBpp0099810	protein_coding

Gene summary [help](#) [Splice variants »](#)

Name [ey](#) (FlyBaseName gene)

Gene type Known protein coding

Prediction Method Feature imported from [FlyBase](#).

Transcripts

Configuring the display

Tip: use the "Configure this page" link on the left to show additional data in this region.

Bases de données biomoléculaires

Génomique comparative

Integr8 - access to complete genomes and proteomes

<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

Curated versions of EMBL entries for complete genome sequences

IPI

A top-level guide to the main databases that describing higher eukaryotic proteomes

EBI > Databases > Integr8

Integr8 : Access to complete genomes and proteomes

Search for species

Go!

Search for gene/protein

Go!

e.g. "coli", "9606"

e.g. "ras1", "P22981", "GO:0007257", "GO:mitosis"

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

Current Status

Focal Point

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.

bacteria = 712

eukaryota = 66

archaea = 53

History

Latest species

GAS top 10

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

Integr8 - genome summaries

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



EMBL-EBI All Databases Enter Text Here

Databases Tools EBI Groups Training Industry About Us Help Site Index

Home
local help

Integr8 News
Focal Point archive
Latest Species

Browse Species

O.sativa Nipponbare
Literature
Genome Statistics
Proteome Analysis
Downloads
Taxonomy

Inquisitor status

BioMart

Proteomes and Genomes FASTA

About Integr8
Publications
Integr8 Web service

Genome Reviews
Curated versions of EMBL entries for complete genome sequences

IPI
A top-level guide to the main databases that describing higher eukaryotic proteomes

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	3865	—	43261740	1194.665	43.8%	11%	3846
Chromosome 2	3070	—	35954743	1200.836	43.3%	10%	3064
Chromosome 3	3362	—	36192742	1190.964	43.7%	11%	3345
Chromosome 4	2386	—	35498469	1211.117	44.2%	8%	2385
Chromosome 5	2159	—	29737217	1158.583	44%	8%	2153
Chromosome 6	2145	—	30731886	1208.862	43.6%	8%	2143
Chromosome 7	2090	—	29644043	1192.207	43.5%	8%	2082
Chromosome 8	1801	—	28434780	1193.421	43.4%	8%	1797
Chromosome 9	1464	—	22696651	1177.715	43.5%	8%	1459
Chromosome 10	1436	—	22685906	1205.82	43.6%	8%	1434
Chromosome 11	1581	—	28386948	1261.507	42.9%	7%	1575
Chromosome 12	1501	—	27566993	1201.039	43%	6%	1495
Mitochondrion	53	—	490520	824.268	43.9%	9%	53
Chloroplast	88	○	134551	709.524	39%	33%	86
	26838						

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 **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.rerio	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 **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	*


Bases de données biomoléculaires

Domaines protéiques

[ExPASy Home page](#) [Site Map](#) [Search ExPASy](#) [Contact us](#) [Swiss-Prot](#) [ENZYME](#)

Search for

[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


☐ add wildcard ^{***}

Browse:

- [by documentation entry](#)
- [by ProRule description](#)
- [by taxonomic scope](#)
- [by number of positive hit](#)


SRS - Sequence Retrieval System

PROSITE tools

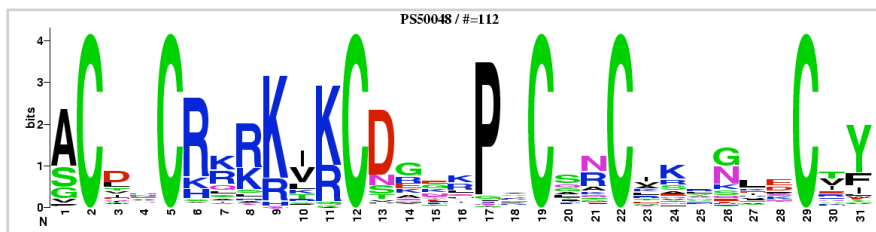
Scan a sequence against PROSITE patterns and profiles - quick scan
(Output includes graphical view and feature detection)

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

☒ exclude [patterns with a high probability of occurrence](#)

- **ScanProsite** - advanced scan
- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** ^{new} - allows to generate custom domain figures.





- La figure ci-dessous représente un profil de domaine conservé (graphique dit « **logo** »). Il indique la conservation des acides aminés à chaque position sein d'un domaine protéique.
- Le logo est obtenu à partir de l'alignement d'une série de séquences protéiques.
- Illustration
 - Profil "Prosite" pour le domaine de liaison à l'ADN "Zn(2)-C6, caractéristique des Fungi (ZN2_CY6_FUNGAL_2, PS50048).
 - A noter
 - Les 6 cystéines très conservées, caractéristiques de ce domaine.
 - La présence d'un espace de taille variable entre les deux moitiés du domaine.



```

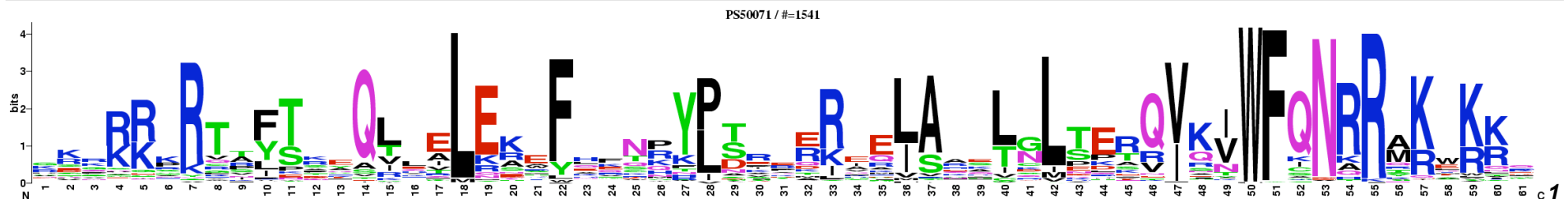
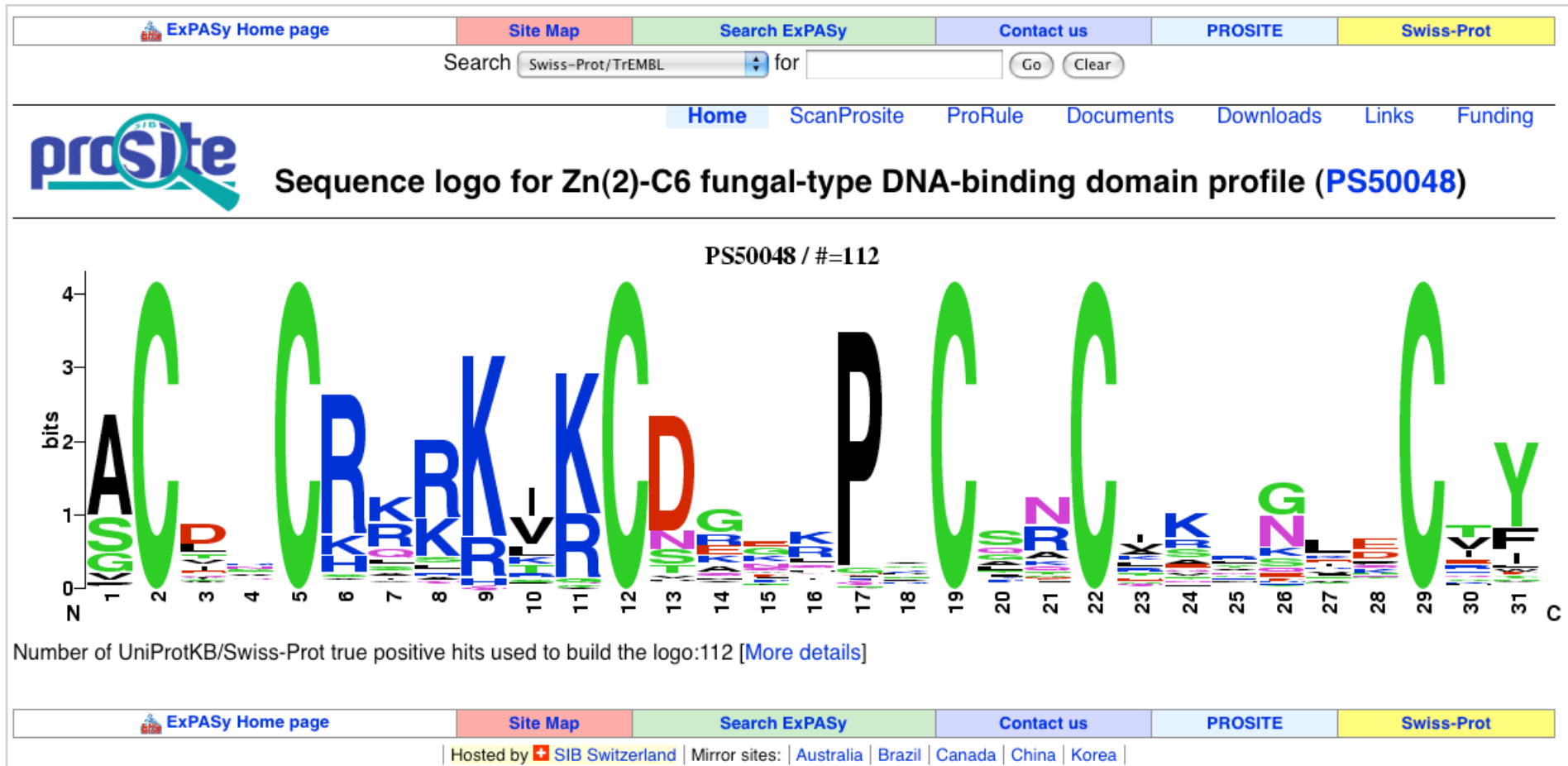
ACE2_TRIPE/6-38
ACR2_NEUCR/21-49
ACU15_NEUCR/23-53
AFLR_ASPFL/28-58
AFLR_ASPFA/28-58
AFLR_EMENI/27-57
ALCR_EMENI/11-51
AMDR_ASPFU/26-59
AMDR_ASPOR/25-58
AMDR_EMENI/19-52
ARGR2_YEAST/20-48
ARO80_YEAST/24-60
ATG2_PICPA/631-660
CAT8_YEAST/69-99
CBF3B_YEAST/13-44
CHA4_YEAST/43-72
CTF1A_FUSSO/60-92
CTF1B_FUSSO/52-83
CZF1_CANAL/317-347
DAL81_YEAST/149-181
ECM22_YEAST/43-73
FCR1_CANAL/25-54
FLUF_NEUCR/10-39
GAL4_YEAST/10-40
GRT1_SCHPO/13-42
HAL9_YEAST/135-168
HAP1_YEAST/63-95
LAC9_KLULA/94-124
LEUR_YEAST/36-69
LYS14_YEAST/158-188
MAL13_YEAST/12-41
MAL33_YEAST/7-36
MAL63_YEAST/7-36
MOC3_SCHPO/35-63
NIRA_EMENI/41-72
NIT4_NEUCR/52-83
OAF1_YEAST/65-95
PDR1_YEAST/45-74
PDR3_YEAST/14-43
PDR8_YEAST/30-61
PIP2_YEAST/24-54
PPR1_YEAST/33-63
PRIB_LENED/19-52
PRO1_NEUCR/54-84
PUT3_YEAST/33-62

ACDRCHDKKLRCPRI.sGSP.....CCSRCAKANV...ACVF
ACYNCHRRKLRCDK...SLP.....ACLKCSINGE...EC--
ACDRCRSKLRCDG...IRP.....CCSQCANVGF...ECKT
SCTSCASSVCTK...EKP.....ACARCIERGL...ACQY
SCTSCASSVCTK...EKP.....ACARCIERGL...ACQY
SCISCSRSVCKNK...EKP.....TCSRVRRL...PCEY
SCDPCRKGRLCDAP.eNRNeanengwvSCSNCKRWNK...DCTF
ACVHCHRRVRCDArivGLP.....-CSNCRSSGKt...DORI
ACIHCHRRVRCDArivGLP.....-CSNCRSAGKa...DORI
ACVHCHRRVRCDArivGLP.....-CSNCRSAGKt...DQOI
GCWTCRGRVRCDL...RHP.....HCQRCEKSNL...PC--
ACISCSRSVRCDLgvpDNPd....pPCARCKRELK...KCIF
GCLTCRKRQVRCDE...RKP.....FCLNCEKSEQ...KCT-
ACDRCRSKLRCDG...KRP.....QSCQAAVGF...ECRI
PCSVCTRRVRCDR...MIP.....-CGNCRKRGQd.sECMK
ACQNCRRRRRCNM...EKP.....-CSNCKIFRT...ECVF
ACETCHARVRCDAasIGVP.....-CTNCVAFQI...ECRI
ACVSCRARVRCDVv.eGAP.....-CGNCRWDNV...ECVV
GCLTCRQRKRCCE...TRP.....RCTECLRLRL...NCTW
SCNQCRLLKRCNYf.pDLG.....NLECETSRT...KCTF
GCNCKRRVRCDE...GKP.....FGKKCTNMKL...DQVY
ACDSCKRKLRCDG...KKP.....-CNRCTLDNK...ICVF
ACLVCRKKLRCDG...QMP.....-CRRCRSRGE...ECAY
ACDICRLKLRCSK...EKP.....KCAKCLKNW...ECRY
ACENCRKRVRCSG...GDV.....-CFECQKYNE...NCVY
ACDHCRRKLRCDV...DQQt....kKCSNCKIFQL...PCTF
SCTICRKRVRCDK...LRP.....HCQQCTKTGva.hLCHY
ACDACRKKLRCSK...TVP.....TCTNCLKYNL...DQVY
ACVECRQQSLCDAh.eRApe.....PCTKCAKKNV...PCIL
GCSECKRRVRCDE...TKP.....TCWQCARLNR...QCVY
ACDCCRIRVRCDG...KRP.....-GSSLQNSL...DCTY
ACDYCRVRVRCDG...KKP.....-CSRCIEHNF...DCTY
SCDCCRVRVRCDR...NKP.....-CNRCIQNL...NCTY
GCLTCRRRIRCDE...TKP.....FCLNCTKTNR...EC--
ACIACRRRSKCDG...NLP.....SCAACSSVYHt...TCVY
ACIACRRRSKCDG...ALP.....SCAACSVYgt...ECIY
VCQACWKSRCDR...EKP.....ECGRCKVHGL...KQVY
ACDNCRRKLRCSG...KFP.....-CASCEIYSC...ECTF
ACVNCRRKLRCTG...KYP.....-CTNCISYDC...TCVF
SCAFCRKRLRCSQ...ARP.....MCQQCVIRKLp...QCVY
VCQACRKAIRCDQ...EKP.....RCGRCTKQNL...FCIY
ACKRCRLKLRCDQ...EFP.....SKRCACKLEV...PCVS
ACTTCRAAMRCVGa...EDGq....rCQRCKRANV...QCIF
GCTTCRLRKRCD...GSP.....MCTACKHLGL...CCEY
ACLSCKRKRIRCPG...GNP.....-CQKCVTSNA...ICEY
    
```

ExPASy Home page		Site Map	Search ExPASy	Contact us	PROSITE
Hosted by  SIB Switzerland		Mirror sites: Australia Brazil Canada China Korea			
Search		<input type="text" value="PROSITE"/>	<input type="button" value="Go"/>	<input type="button" value="Clear"/>	
		Home	ScanProsite	ProRule	Documents
			Downloads	Links	Funding
Entry: PS50048					
General information about the entry					
Entry name	ZN2_CY6_FUNGAL_2				
Accession number	PS50048				
Entry type	MATRIX				
Date	NOV-1997 (CREATED); NOV-1997 (DATA UPDATE); NOV-2008 (INFO UPDATE).				
PROSITE Documentation	PDOC00378				
Associated ProRule	PRU00227				
Name and characterization of the entry					
Description	Zn(2)-C6 fungal-type DNA-binding domain profile.				
Matrix / Profile	<pre>/GENERAL_SPEC: ALPHABET='ABCDEFGHIJKLMNPQRSTVWYZ'; LENGTH=31; /DISJOINT: DEFINITION=PROTECT; N1=4; N2=28; /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.3; R2=0.011; TEXT='-LogE'; /CUT_OFF: LEVEL=0; SCORE=801; N_SCORE=8.5; MODE=1; TEXT='!'; /CUT_OFF: LEVEL=-1; SCORE=619; N_SCORE=6.5; MODE=1; TEXT='?'; /DEFAULT: D=-20; I=-20; B1=-20; E1=-20; MI=-105; MD=-105; IM=-105; DM=-105; A B C D E F G H I K L M N P Q R S T V W Y Z /I: B1=0; B1=-105; BD=-105; /M: SY='A'; M= 26, -9,-14,-14,-10,-20, 5,-18,-15,-12,-17,-13, -6, -7,-10,-17, 13, 3, -5,-26,-21,-10; /M: SY='C'; M=-10,-20,120,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30; /M: SY='D'; M=-10, 8,-25, 12, -1,-19,-18,-10,-12, -8,-10,-10, 0,-16, -6,-10, -4, -2, -9,-22,-10, -3; /M: SY='H'; M= -6, -4,-16, -9, -3,-11,-18, 1,-13, -6,-12, -7, 1,-15, 1, -1, 0, 0,-12,-25, -6, -2; /M: SY='C'; M=-10,-20,120,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30; /M: SY='R'; M=-15, -8,-27, -8, 0,-21,-19, 4,-29, 25,-21,-10, 0,-18, 8, 50, -7, -9,-19,-20, -8, 0; /M: SY='R'; M= -9, -9,-25,-11, -2,-16,-19, -9,-15, 13,-11, -4, -5,-17, 5, 20, -7, -6,-11,-21, -9, 0; /M: SY='R'; M=-11, -8,-27, -9, 0,-21,-17, -5,-23, 23,-18, -8, -2,-17, 6, 39, -7, -8,-15,-22,-10, 0; /M: SY='K'; M=-12, -2,-30, -2, 8,-28,-20, 2,-30, 40,-27, -9, 1,-12, 11, 33,-10,-11,-21,-21, -7, 8; /M: SY='V'; M= -6,-19,-21,-23,-17, -7,-27,-21, 16,-10, 4, 7,-15,-21,-13, -9, -9, -1, 17,-24, -7,-17; /M: SY='K'; M=-11, -3,-29, -3, 6,-27,-18, -8,-28, 40,-26,-10, -1,-13, 9, 35, -8, -9,-18,-21,-11, 7; /M: SY='C'; M=-10,-20,120,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30; /M: SY='D'; M=-14, 32,-23, 42, 11,-31,-10, -4,-30, -3,-26,-24, 16,-11, -2, -9, 4, -3,-23,-38,-19, 4; /M: SY='R'; M= -4, -8,-26, -9, 1,-17, -2, -8,-21, 4,-15, -8, -5,-16, -1, 6, -5, -9,-16,-19,-10, -1; /I: II=-8; MI=0; MD=-20; DM=-20; IM=0; I=0,0,-30,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0; /M: SY='E'; M= -5, 3,-25, 2, 5,-23, -7, -7,-19, 3,-16,-10, 3,-12, 2, 0, -2, -6,-17,-26,-15, 3; /M: SY='R'; M=-12, -9,-28,-10, 0,-12,-22, -7,-15, 12,-10, -5, -5,-15, 2, 17,-11, -8,-14,-15, -5, 0; /M: SY='P'; M=-10,-16,-37, -8, -1,-27,-15,-16,-21,-10,-29,-19,-15, 73,-10,-18, -8,-10,-29,-29,-27,-10; /I: II=-8; MI=0; MD=-10; DM=-10; IM=0; I=0,0,-30,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0; /M: SY='S'; M= -2, -3,-14, -6, -3, -9,-11, -2, -7, -1,-10, -4, 0, -6, -1, -1, 1, 0, -5,-17, -6, -3; D=-3; /I: DM=-10; /M: SY='C'; M=-10,-20,120,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30; /M: SY='S'; M= 1, -7,-20,-10, -4,-14, -7,-11,-14, -6,-14, -9, -3,-15, 2, -5, 7, 3,-11,-21,-10, -1;</pre>				

Prosite – Exemple de logo

<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:

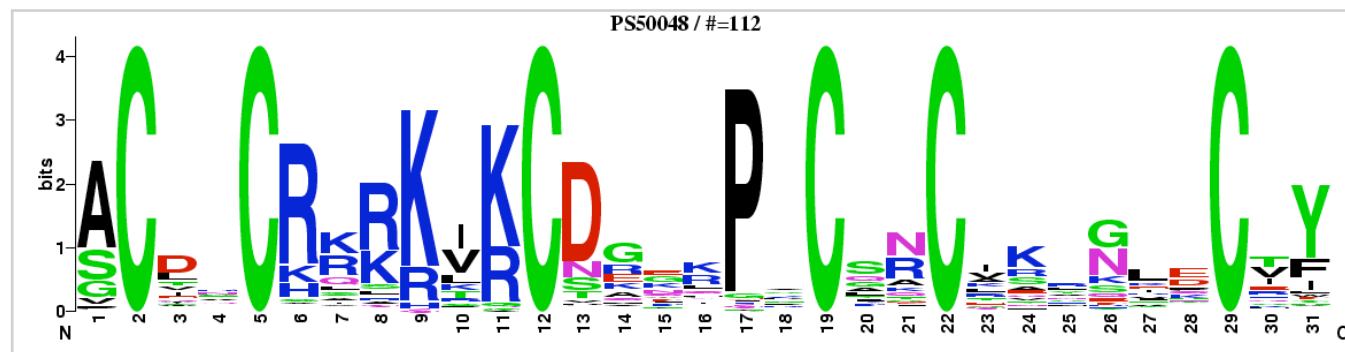
human ultra high-sulfur keratin.

- 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)



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#)



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](#)

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:1557122](#). A wide range of proteins are known to contain this domain. These include the proteins involved in arginine, proline, pyrimidine, quinate, 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	transcription factor activity (GO:0003700)
	regulation of transcription, DNA-dependent (GO:0006355)

Internal database links

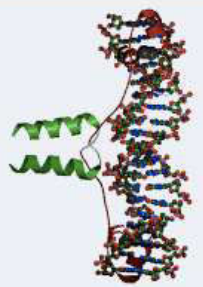
SCOOP:	EndIII_4Fe-2S
--------	-------------------------------

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

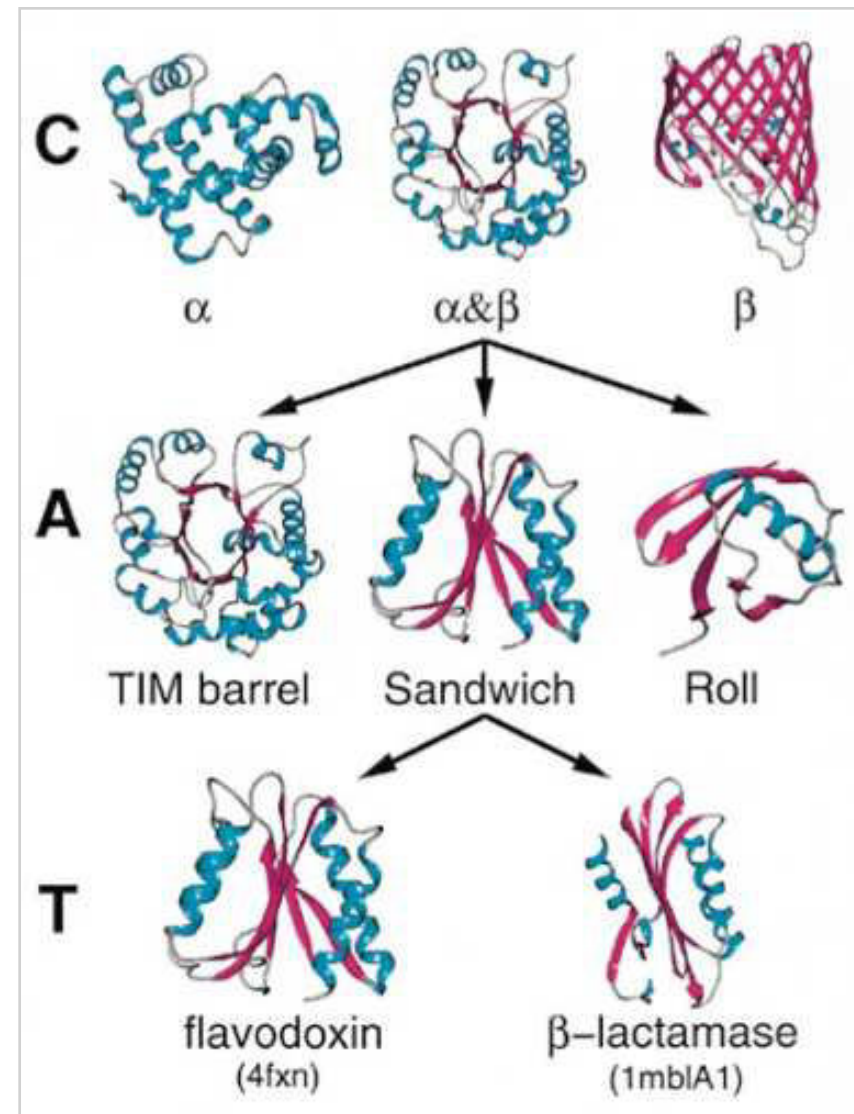


CATH – Classification de structures protéiques

<http://www.cathdb.info/>



- CATH est une classification hiérarchique des domaines observés dans les structures tri-dimensionnelles de protéines.
- Les domaines sont classés selon 4 niveaux hiérarchiques
 - Classe (C),
 - Architecture (A),
 - Topologie (T)
 - Homologie (H).
- Les limites des classes et les assignations des protéines aux classes reposent sur un processus semi-automatique (algorithme + révision humaine).
- Les critères d'élaboration des classes combinent des méthodes informatiques, statistiques, une revue de la littérature et une analyse par des experts.



- Orengo et al. The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res (1999) vol. 27 (1) pp. 275-9
- Cuff et al. The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res (2008) pp.

- Une base de données de familles protéiques, domaines, répétitions et sites au sein desquels on peut identifier des caractéristiques qui peuvent ensuite être appliquées pour analyser de nouvelles séquences protéiques.

- InterPro:Home
 - Advanced Search
 - InterProScan
 - Databases
 - Documentation
 - Release Notes
 - User Manual
 - FAQ
 - Tutorial
 - Example Entry
 - Project Outline
 - People
 - Database Contributors
 - Publications
 - Web Services
 - FTP site
 - Protein of the month
 - Fatty Acid Synthase

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, GENE3D and SUPERFAMILY.
- PROSITE pattern matches** are now evaluated to either TRUE (T) or UNKNOWN (?) 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 domain(s).
- New relationship rules will be introduced that will affect how different entries are related to one another. **Parent/Child** 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.
 - Parent/Child** 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

Signatures

Database	ID	Name	Proteins
PRINTS	PR00025	ANTENNAPEDIA	510
PROSITE pattern	PS00032	ANTENNAPEDIA	973

InterPro Relationships

Parent IPR001356 Homeobox

GO Term annotation

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

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

InterPro annotation

Abstract	<p>The homeobox is a 60-residue motif first identified in a number of <i>Drosophila</i> 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.</p>
	<p>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 <i>Drosophila</i>, 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 <i>Drosophila</i> 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.</p>
	<p>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 D8.</p>
	<p><i>Caenorhabditis elegans</i> lin-39 and mab-5 are also members of the 'antennapedia' subfamily.</p>

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: [a.4.1.1](#)
CATH: [1.10.10.60.20](#), [1.10.10.60.4](#)

Database links

MSDsite: [PS00032](#)
PROSITE doc: [PDOC00032](#)
Blocks: [IPB001827](#)

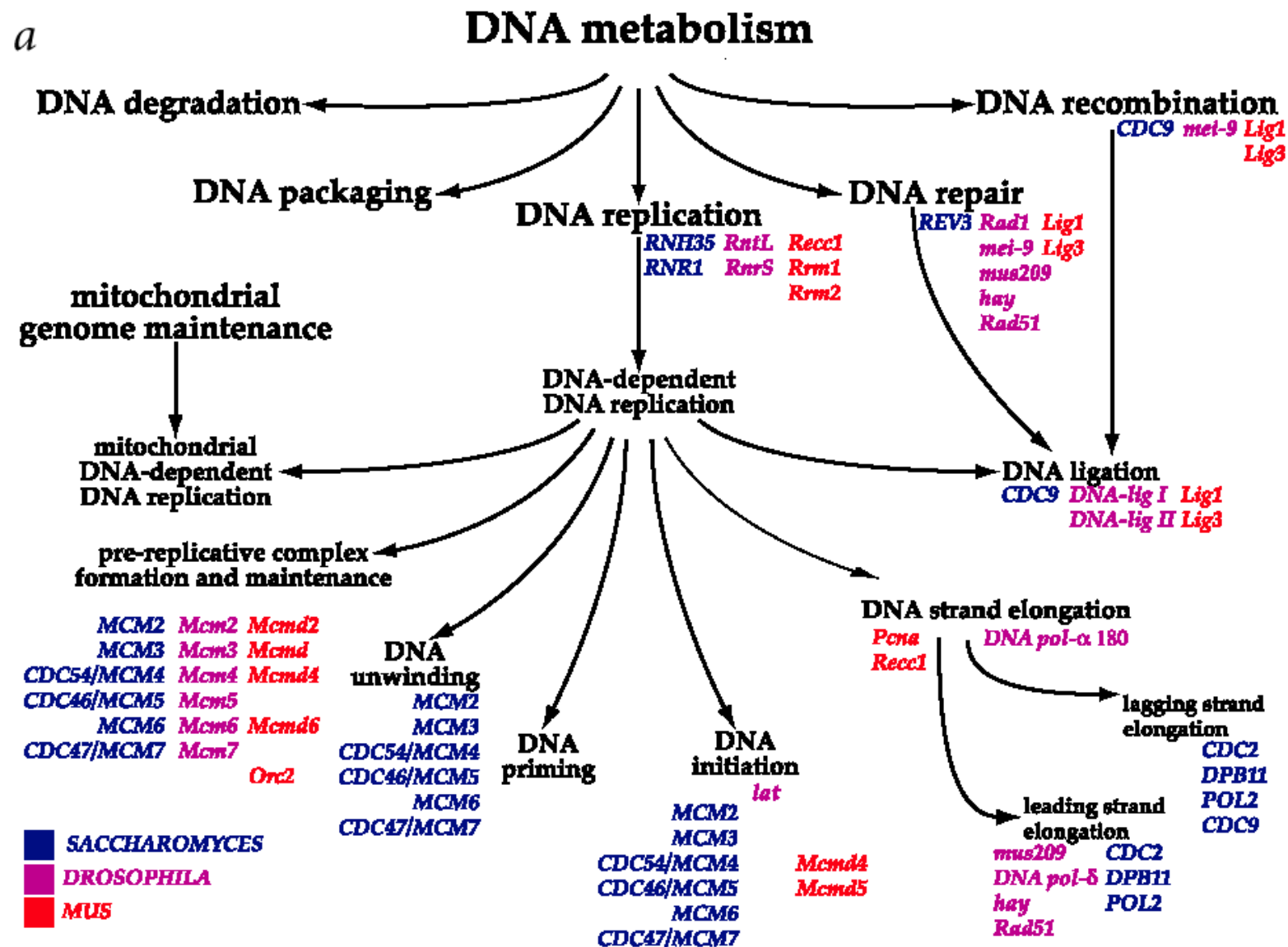
La base de données “Gene ontology” (GO)

Ontologie – définition générale

- Ontologie: partie de la métaphysique qui s'intéresse à l'être en tant qu'être, indépendamment de ses déterminations particulières
- *Le Petit Robert - dictionnaire alphabétique et analogique de la langue française. 1993*

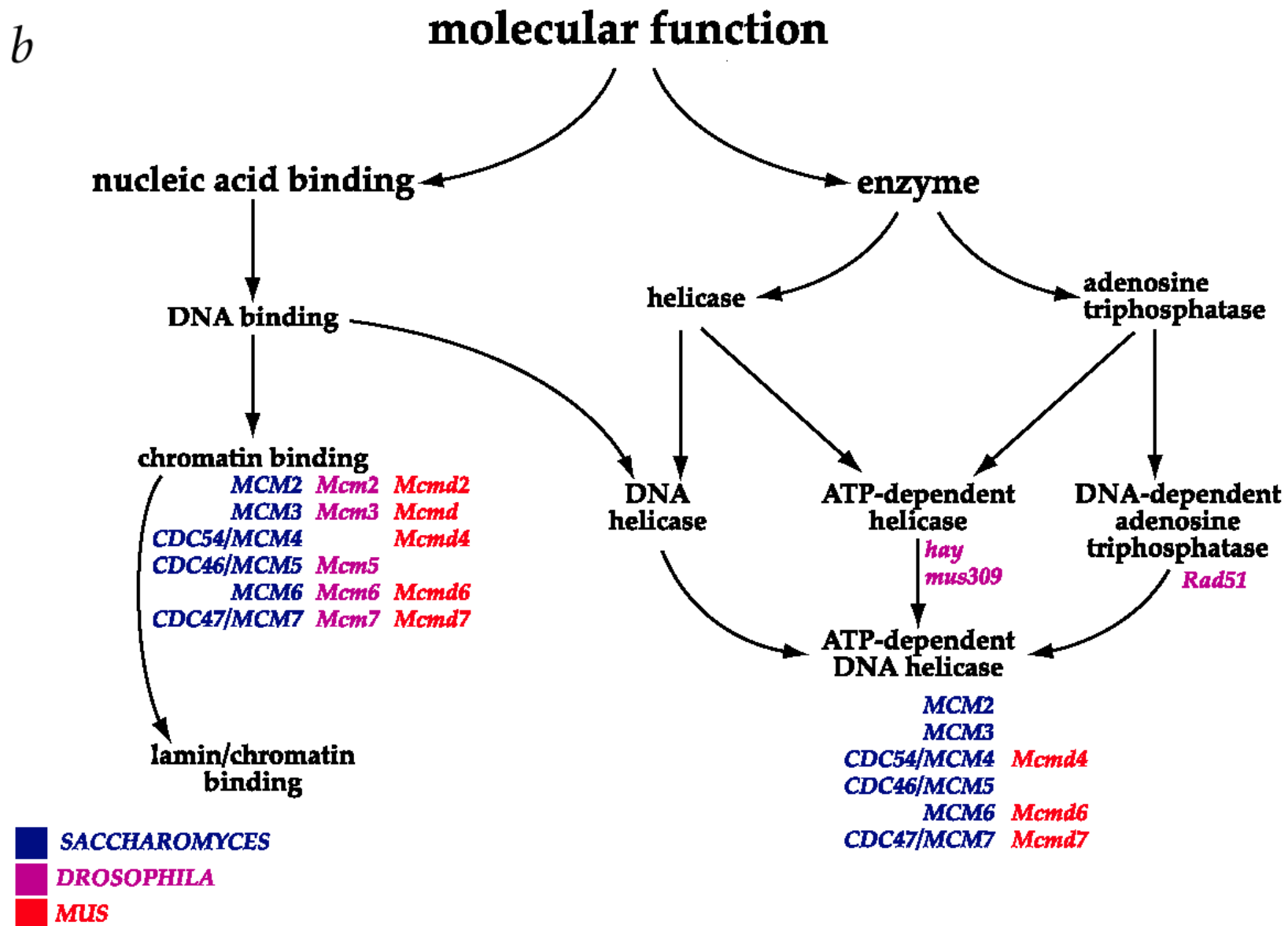
- Les bio-ontologies ne constituent pas une « ontologie » au sens philosophique du terme, elles se rapportent à un sens dérivé en informatique: classification des concepts liés à un champ disciplinaire.
- Les bio-ontologies visent à répondre au problème d'inconsistance entre annotations.
- Pour y répondre, on met en place
 - Un vocabulaire contrôlé
 - On utilise toujours le même mot pour désigner le même concept.
 - Les listes de synonymes permettent d'établir les correspondances.
 - Classification hiérarchique entre les termes de ce vocabulaire contrôlé.
- La « Gene ontology » établit une classification des gènes et protéines selon trois critères complémentaires:
 - Fonction moléculaire (ex: aspartokinase, transporteur de glucose, ...).
 - Processus biologique (ex: biosynthèse de la méthionine, réplication, ...).
 - Composante cellulaire (ex: membrane mitochondriale, noyau, ...).

Gene ontology: exemples de processus biologiques

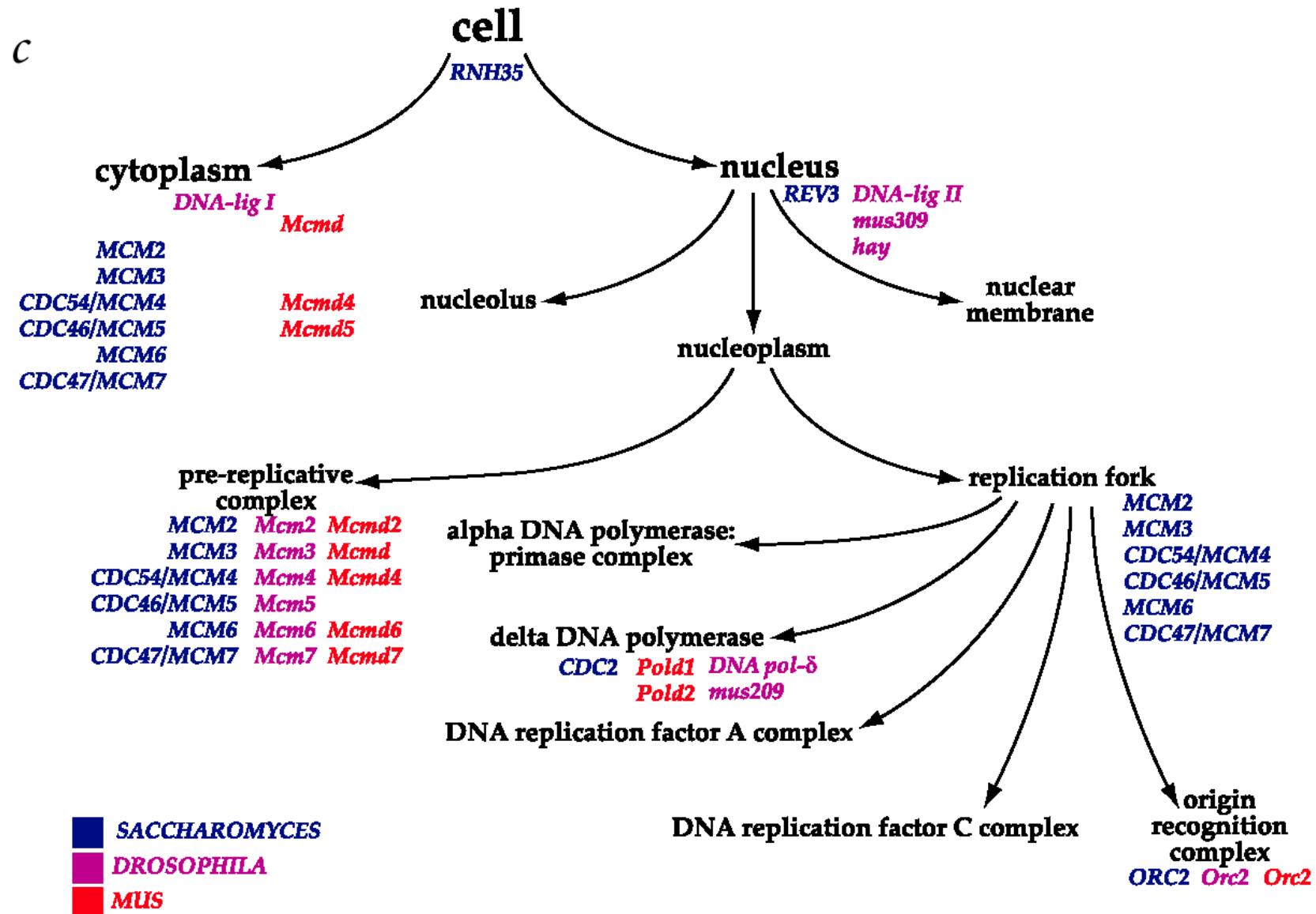


Gene ontology: exemples de fonctions moléculaires

b



Gene ontology: exemples de composantes cellulaires





the Gene Ontology

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

Statut des annotations de GO (NAR DB issue 2006)

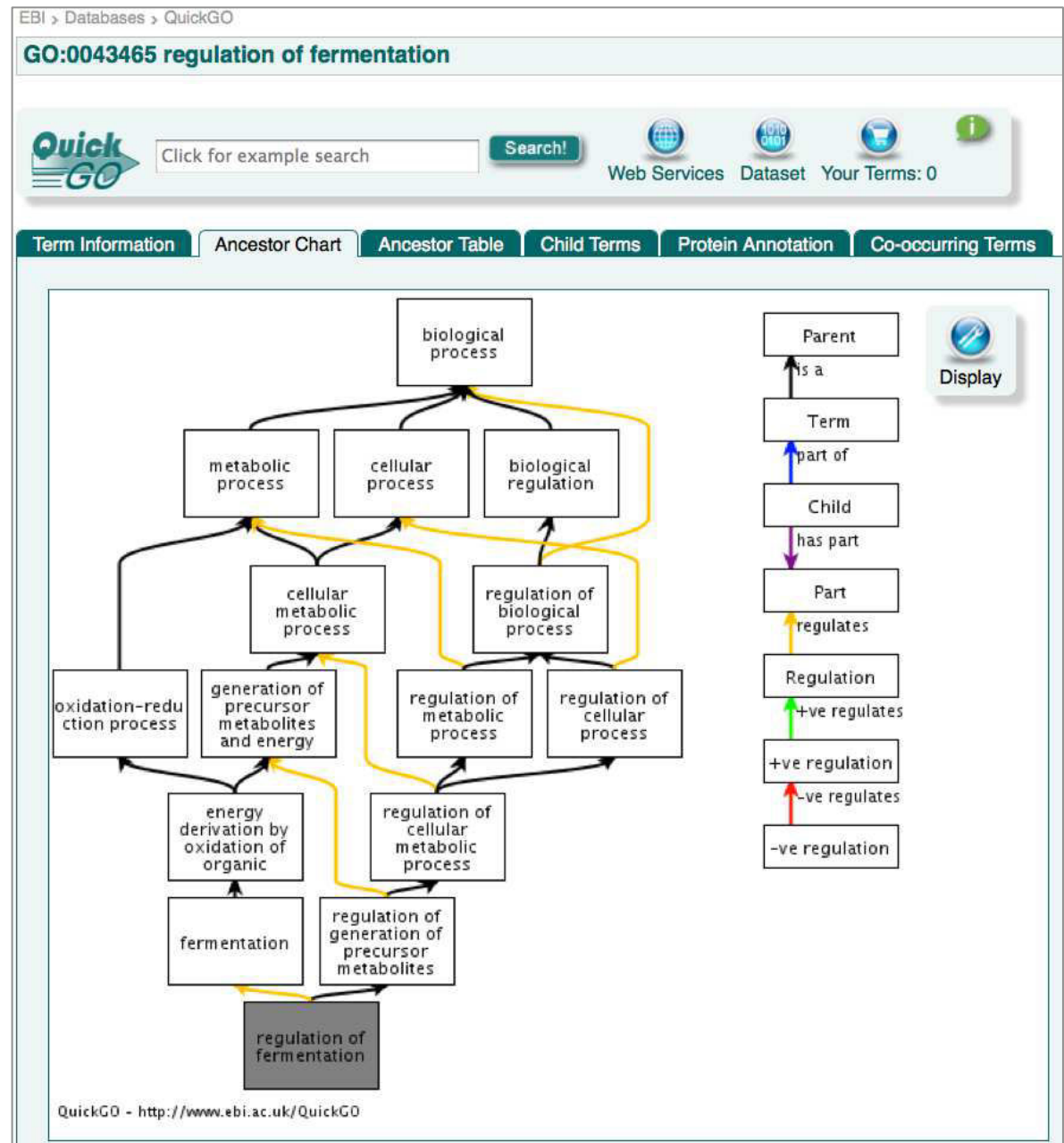
- Nombre de termes définis
 - Biological process 9,805
 - Molecular function 7,076
 - Cellular component 1,574
 - Sequence Ontology 963

- Génomes annotés 30
 - Ceci ne prend pas en compte les annotations d'Uniprot, qui comptaient à l'époque protéomes.
 - Le nombre de génomes annotés augmente de façon exponentielle avec le temps. Il existe aujourd'hui (2013) des milliers de génomes annotés.

- Annotations de produits de gènes
 - Total 1,618,739
 - Automatique (informatique) 1,460,632
 - Curation "manuelle" (révision par des humains) 158,107

QuickGO (<http://www.ebi.ac.uk/QuickGO/>)

- Site Web
<http://www.ebi.ac.uk/QuickGO/>
- Une interface conviviale pour la Gene Ontology.
- Affichage graphique de la hiérarchie entre les termes.
- Navigation aisée entre les classes.



Remarques concernant la « bio-ontologie »

- Améliorations par rapport aux annotations en texte libre
 - Vocabulaire contrôlé (termes définis + synonymes)
 - Relations hiérarchiques entre les concepts
- Aucun rapport avec le concept philosophique d'ontologie.
 - Une « bio-ontologie » n'est rien de plus qu'une classification des termes d'un vocabulaire contrôlé.
- Simplification de la réalité
 - Il existe des critères multiples de classification. Par exemple:
 - Sous-types de compartiments (une membrane plasmique est une membrane)
 - Localisation des compartiment: le noyau se situe à l'intérieur de l'enveloppe nucléaire, qui se situe dans le cytoplasme, qui est à l'intérieur de la membrane plasmique, ...)
- Représentation très incomplète
 - Les ontologies consistent à annoter les objets individuels, mais n'indiquent pas les relations entre eux. Ex:
 - Relation facteur transcriptionnel -> gène cible
 - Interactions protéiques
 - Interconnexions entre réactions d'une voie métabolique

Comment définir la fonction biologique ?

■ A general definition

- Fonction: action, rôle caractéristique d'un élément, d'un organe, dans un ensemble (souvent opposé à structure). Source: Le Petit Robert - dictionnaire alphabétique et analogique de la langue française. 1982.

■ Représentation de la fonction dans la « gene ontology »

- Dans la « gene ontology », on a une vision très fragmentaire de la fonction, car la structure même des données sépare les éléments complémentaires.
- Pour comprendre la fonction, il faut établir le lien entre l'action (activité moléculaire) et le contexte au sein duquel cette action prend place (processus biologique).
- Ceci est nécessaire pour pouvoir traiter la multifonctionnalité:
 - Une même activité peut jouer différents rôles dans différents processus
 - Ex: le gène *scure* de la drosophile code pour un facteur transcriptionnel (activité) qui, selon le stade de développement et la localisation tissulaire, est impliqué dans différents processus: détermination du sexe, des précurseurs neuraux, ou des précurseurs de tubules de Malpighi.
 - Une même protéine peut jouer différents rôles dans le même processus.
 - Ex: aspartokinase PutA chez *Escherichia coli*, contient 2 domaines enzymatiques + un domaine de liaison à l'ADN -> 3 activités moléculaires impliquées à différents niveaux du même processus (utilisation de la proline).

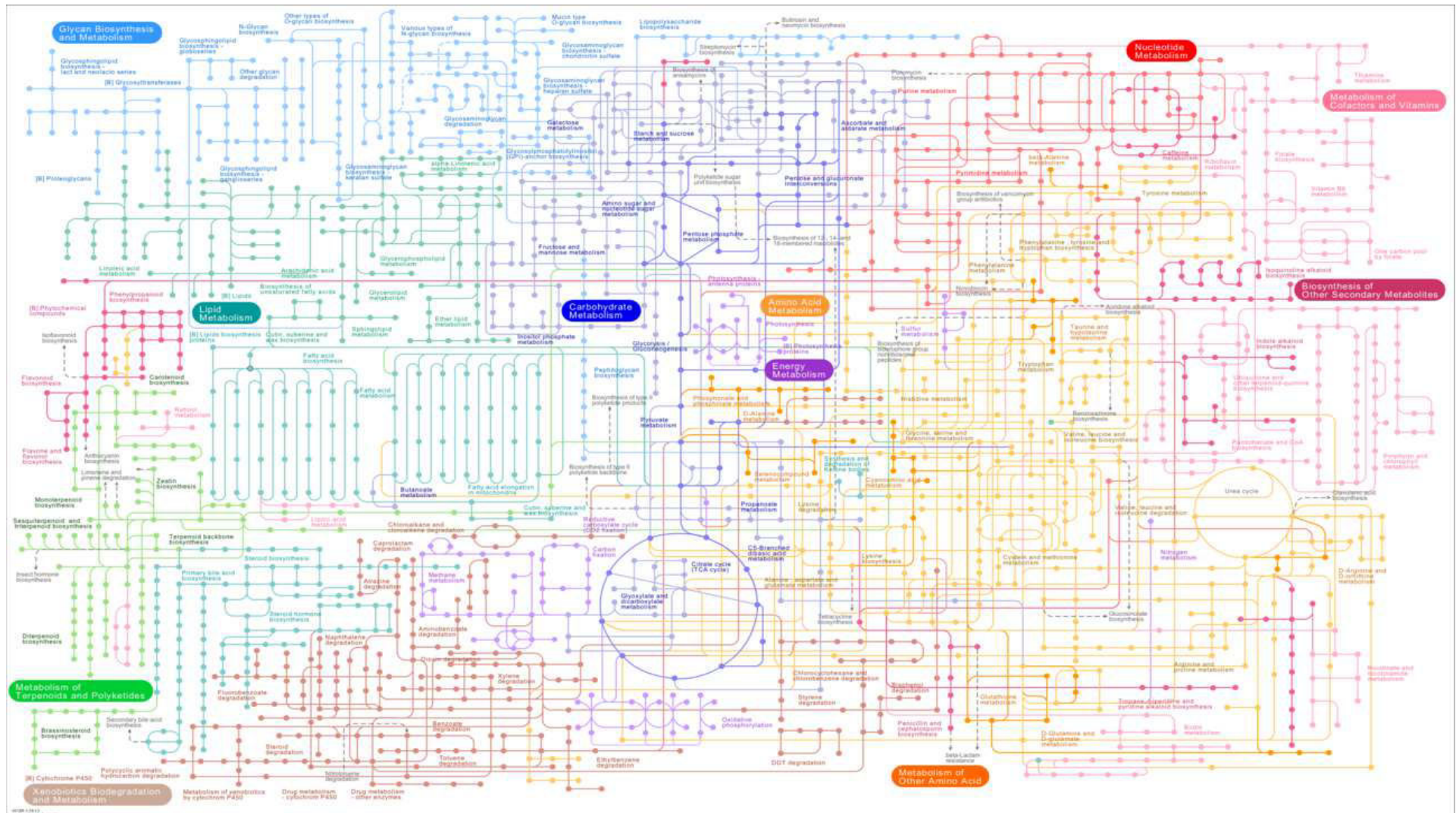
Bases de données biomoléculaires

***Petites molécules,
réactions biochimiques
et voies métaboliques***

LIGAND – réactions métaboliques et petites molécules

KEGG - Kyoto Encyclopaedia of Genes and Genomes

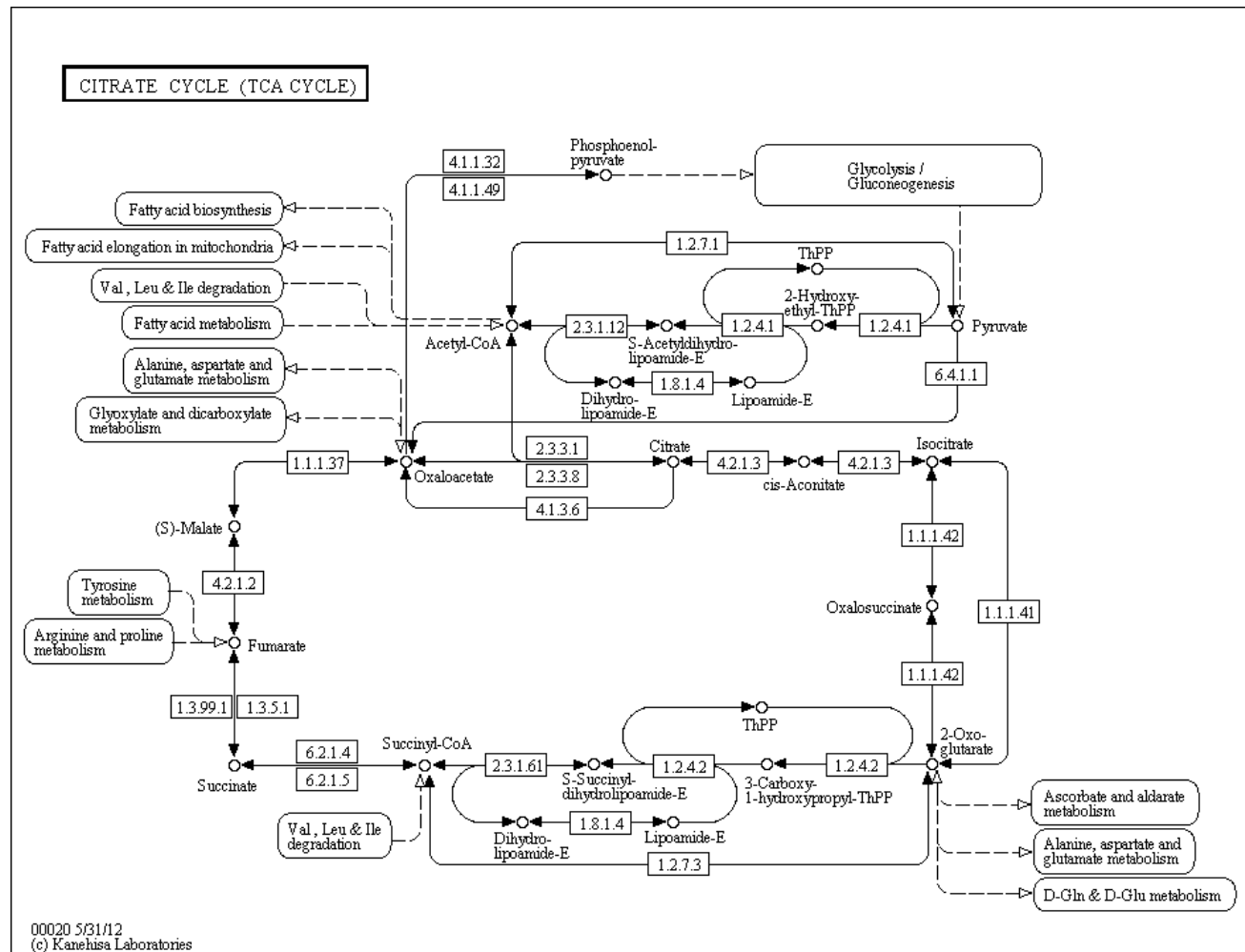
- La “carte globale” donne une vue d'ensemble de la complexité du métabolisme. Chaque point représente une molécule, chaque ligne une réaction métabolique.



- KEGG global pathway map: http://www.genome.jp/kegg-bin/show_pathway?map01100

KEGG - Kyoto Encyclopaedia of Genes and Genomes

- Les cartes métaboliques de KEGG présentent le détail des réactions d'une voie métabolique, en montrant les voies alternatives présentes chez différents organismes.



Ecocyc, BioCyc and Metacyc – Voies métaboliques

***Réseaux d'interactions protéiques
et voies de transduction de signal***

Bases de données biomoléculaires

Bases de données de biopuces

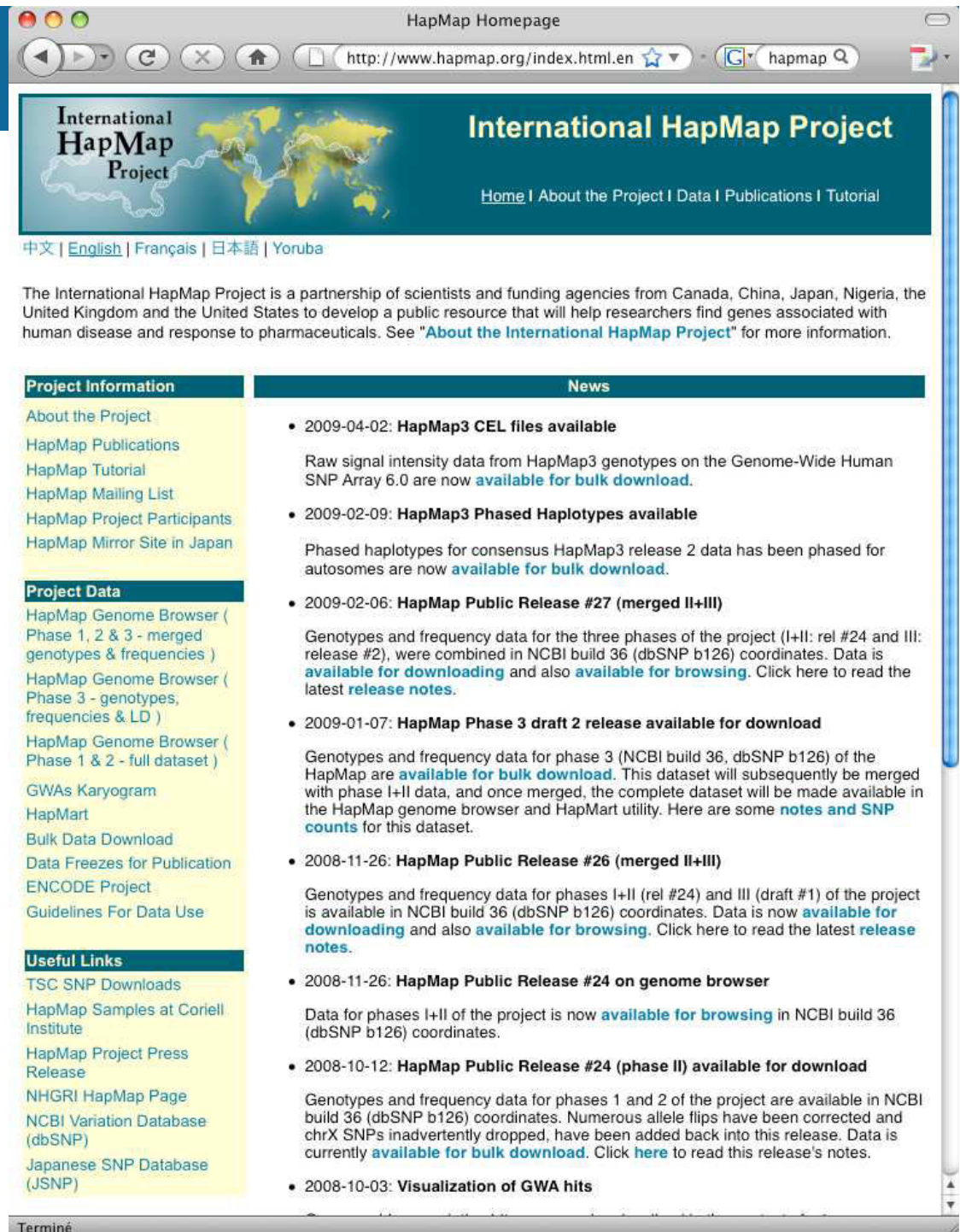
Bases de données biomoléculaires

***Ressources pour l'analyse du
génom humain***

HapMap

<http://www.hapmap.org/>

- HapMap est un projet international pour identifier et cataloguer les similarités et différences génétiques entre humains.
- Cette base de données permet d'étudier les associations entre variations génétiques (SNPs, microsatellites) et
 - maladies.
 - réponse aux médicaments.
- Applications
 - Identification de gènes associés à des maladies.
 - Médecine personnalisée: adaptation des traitements médicaux aux particularités individuelles des patients.



The screenshot shows the HapMap Homepage in a web browser. The browser's address bar displays <http://www.hapmap.org/index.html.en>. The page features a header with the "International HapMap Project" logo and a world map. Below the header, there are navigation links in multiple languages (Chinese, English, Français, 日本語, Yoruba) and a paragraph describing the project as a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom, and the United States. The main content area is divided into two columns: "Project Information" and "News". The "Project Information" column lists links for "About the Project", "HapMap Publications", "HapMap Tutorial", "HapMap Mailing List", "HapMap Project Participants", and "HapMap Mirror Site in Japan". The "Project Data" section lists links for "HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)", "HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)", "HapMap Genome Browser (Phase 1 & 2 - full dataset)", "GWAs Karyogram", "HapMart", "Bulk Data Download", "Data Freezes for Publication", "ENCODE Project", and "Guidelines For Data Use". The "Useful Links" section lists links for "TSC SNP Downloads", "HapMap Samples at Coriell Institute", "HapMap Project Press Release", "NHGRI HapMap Page", "NCBI Variation Database (dbSNP)", and "Japanese SNP Database (JSNP)". The "News" column contains a list of recent updates, including "HapMap3 CEL files available", "HapMap3 Phased Haplotypes available", "HapMap Public Release #27 (merged II+III)", "HapMap Phase 3 draft 2 release available for download", "HapMap Public Release #26 (merged II+III)", "HapMap Public Release #24 on genome browser", "HapMap Public Release #24 (phase II) available for download", and "Visualization of GWA hits".

Project Information

- About the Project
- HapMap Publications
- HapMap Tutorial
- HapMap Mailing List
- HapMap Project Participants
- HapMap Mirror Site in Japan

Project Data

- HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)
- HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)
- HapMap Genome Browser (Phase 1 & 2 - full dataset)
- GWAs Karyogram
- HapMart
- Bulk Data Download
- Data Freezes for Publication
- ENCODE Project
- Guidelines For Data Use

Useful Links

- TSC SNP Downloads
- HapMap Samples at Coriell Institute
- HapMap Project Press Release
- NHGRI HapMap Page
- NCBI Variation Database (dbSNP)
- Japanese SNP Database (JSNP)

News

- 2009-04-02: **HapMap3 CEL files available**
Raw signal intensity data from HapMap3 genotypes on the Genome-Wide Human SNP Array 6.0 are now [available for bulk download](#).
- 2009-02-09: **HapMap3 Phased Haplotypes available**
Phased haplotypes for consensus HapMap3 release 2 data has been phased for autosomes are now [available for bulk download](#).
- 2009-02-06: **HapMap Public Release #27 (merged II+III)**
Genotypes and frequency data for the three phases of the project (I+II: rel #24 and III: release #2), were combined in NCBI build 36 (dbSNP b126) coordinates. Data is [available for downloading](#) and also [available for browsing](#). Click here to read the latest [release notes](#).
- 2009-01-07: **HapMap Phase 3 draft 2 release available for download**
Genotypes and frequency data for phase 3 (NCBI build 36, dbSNP b126) of the HapMap are [available for bulk download](#). This dataset will subsequently be merged with phase I+II data, and once merged, the complete dataset will be made available in the HapMap genome browser and HapMart utility. Here are some [notes and SNP counts](#) for this dataset.
- 2008-11-26: **HapMap Public Release #26 (merged II+III)**
Genotypes and frequency data for phases I+II (rel #24) and III (draft #1) of the project is available in NCBI build 36 (dbSNP b126) coordinates. Data is now [available for downloading](#) and also [available for browsing](#). Click here to read the latest [release notes](#).
- 2008-11-26: **HapMap Public Release #24 on genome browser**
Data for phases I+II of the project is now [available for browsing](#) in NCBI build 36 (dbSNP b126) coordinates.
- 2008-10-12: **HapMap Public Release #24 (phase II) available for download**
Genotypes and frequency data for phases 1 and 2 of the project are available in NCBI build 36 (dbSNP b126) coordinates. Numerous allele flips have been corrected and chrX SNPs inadvertently dropped, have been added back into this release. Data is currently [available for bulk download](#). Click [here](#) to read this release's notes.
- 2008-10-03: **Visualization of GWA hits**