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.

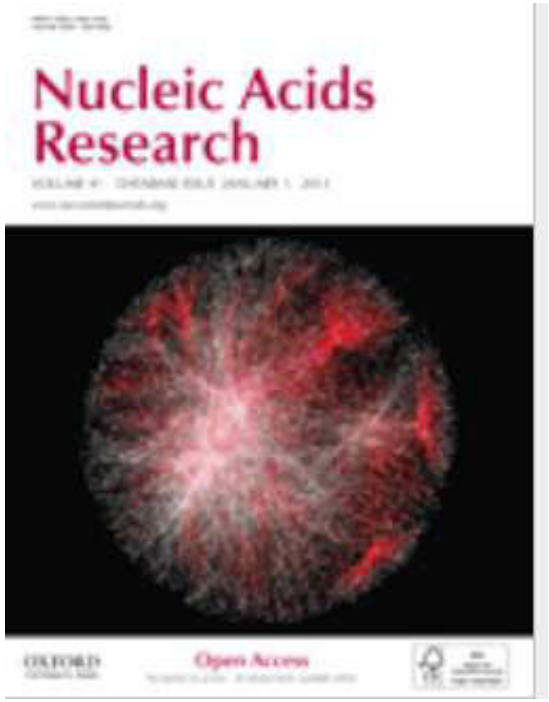
Un tour d'horizon de quelques bases de données biomoléculaires

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, PlasmoDB, …)
- 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.



- Fernandez-Suarez and Galperin, 2013, NAR D1-D7

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)

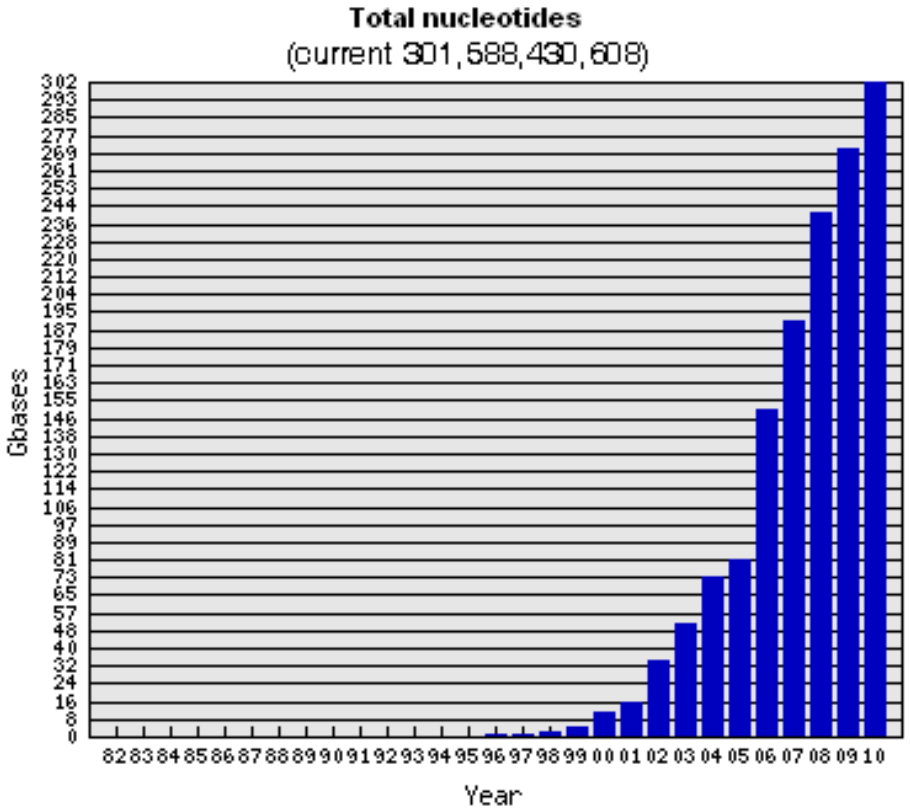
EMBL

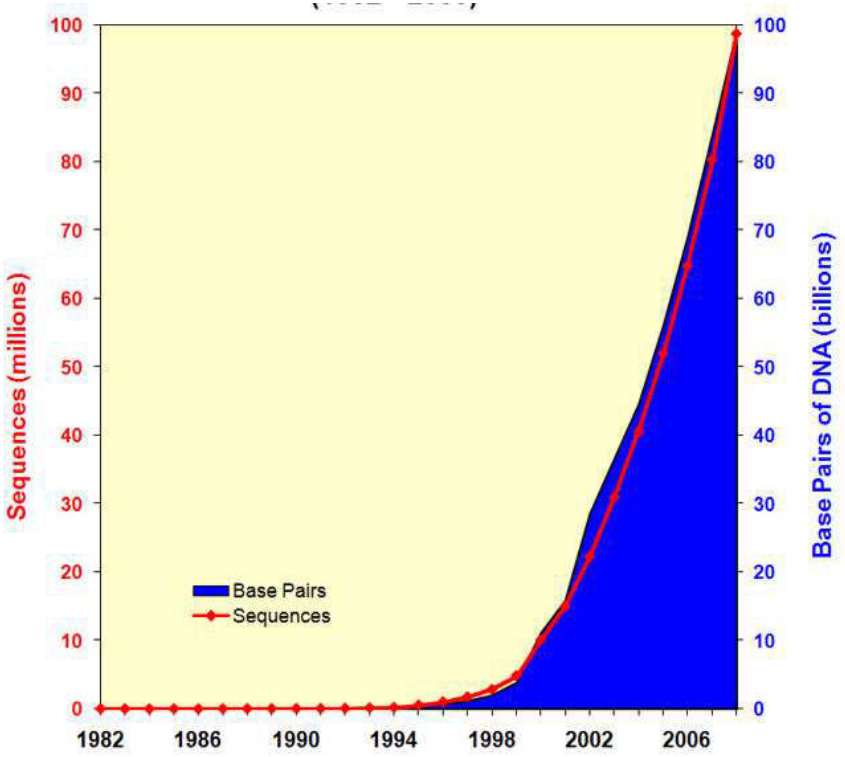
300 Gb (décembre 2010)

Genbank

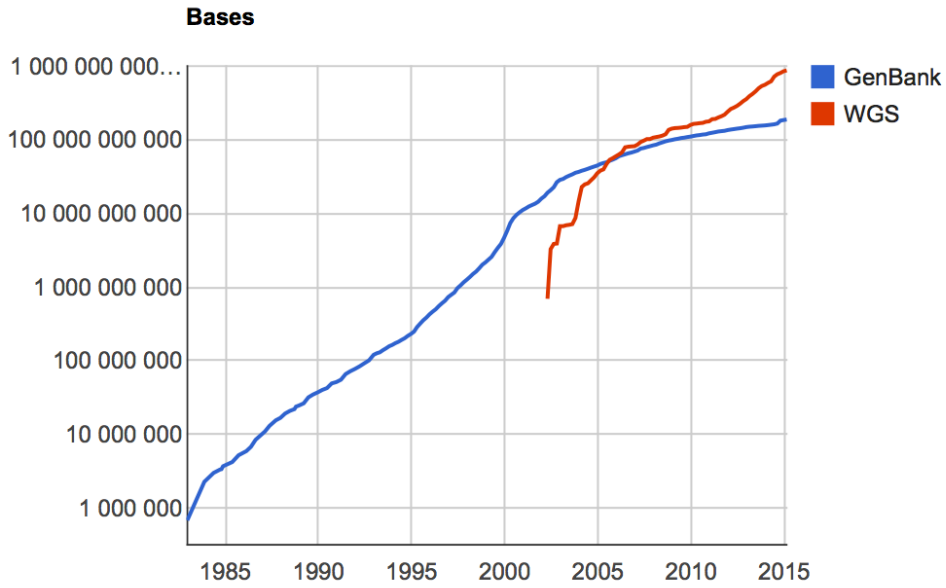
286 Gb (août 2010)

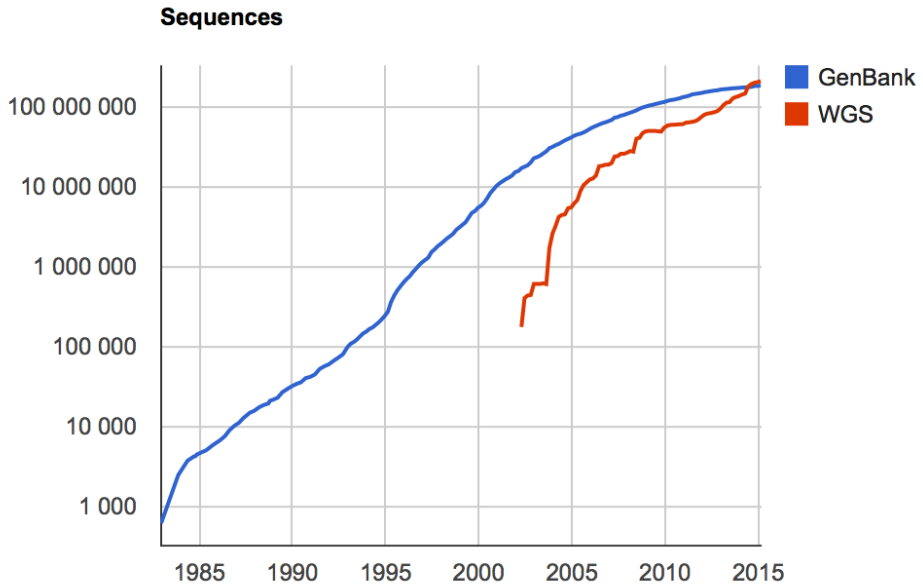
http://www.ebi.ac.uk/embl/Services/DBStats/ http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html





Genbank statistics -2015

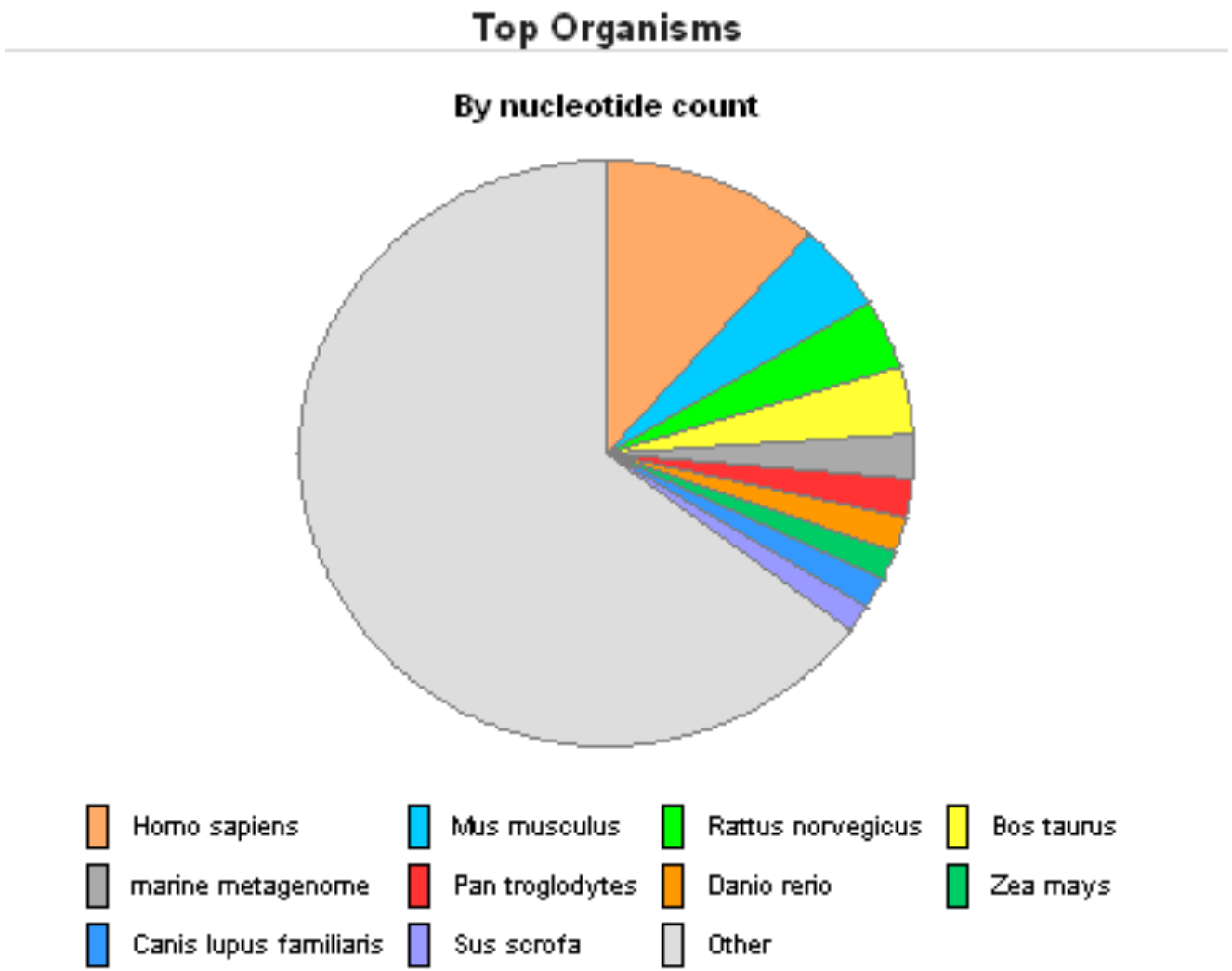




- La taille des données entreposées augmente de façon exponentielle.
- Depuis le début des années 2000, les projets de séquençage génomique constitue une proportion croissante des séquences déposées dans les bases de données.

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

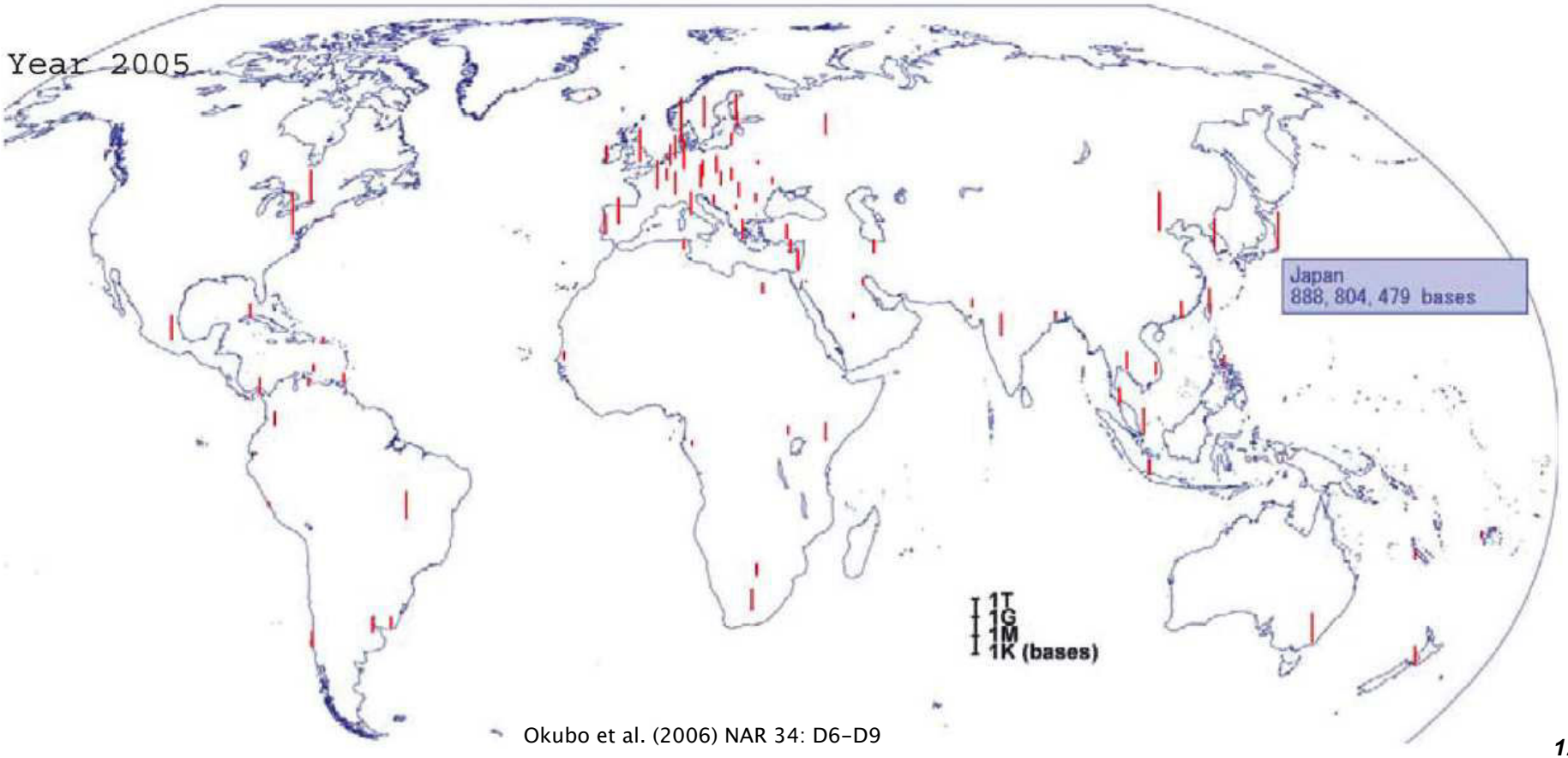
Organismes les plus représentés



EMBL (décembre 2010) http://www.ebi.ac.uk/embl/Services/DBStats/

Bases de données de séquences nucléiques

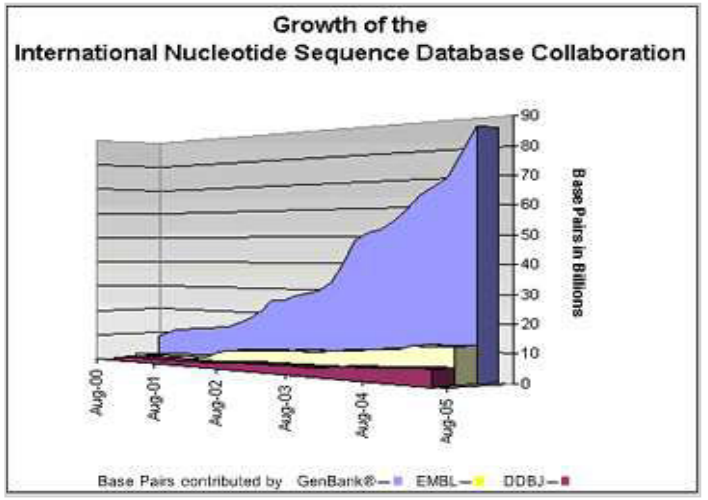
- Avant de publier un article qui décrit une séquence biologique, il est obligatoire de déposer cette séquence dans l'une des 3 principales bases de données de séquences: Genbank (Etats-Unis), EMBL (Europe) ou DDBJ (Japon).
- Les séquences sont automatiquement synchronisées entre les trois bases de données.

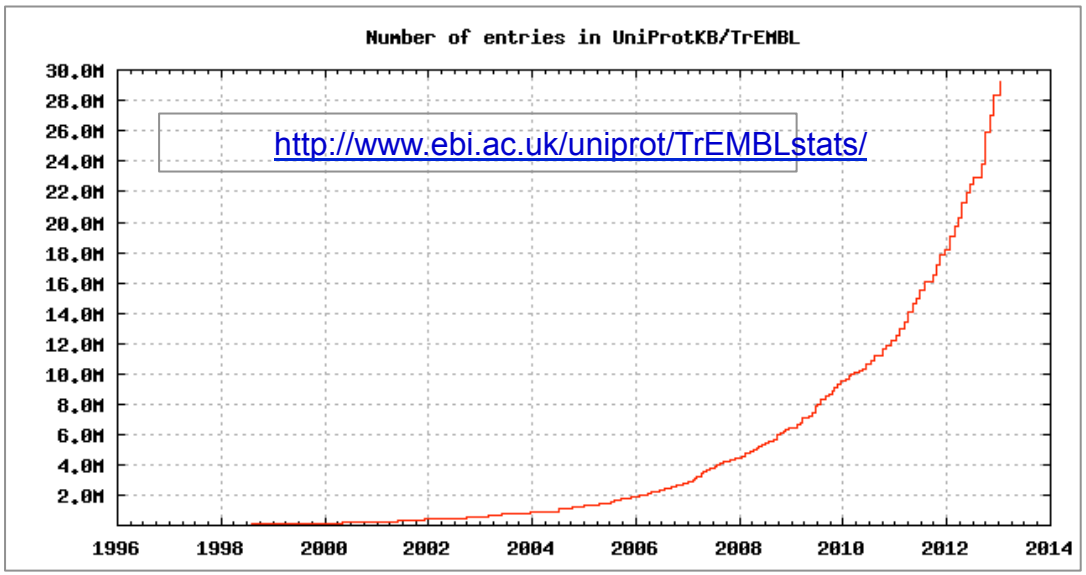


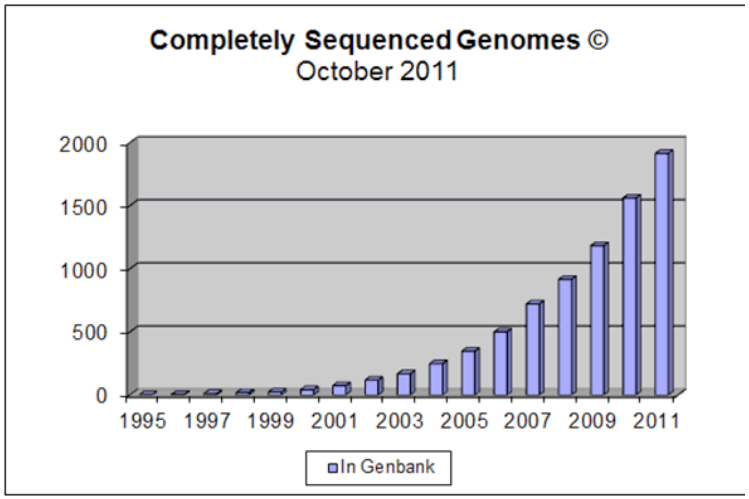
Okubo et al. (2006) NAR 34: D6–D9

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 expontentielle.
- Jusqu'en 2007, cet efet 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 décroit de façon beaucoup plus rapide que celui du stockage, ce qui commence à créer de vrais problème de gestion des données.

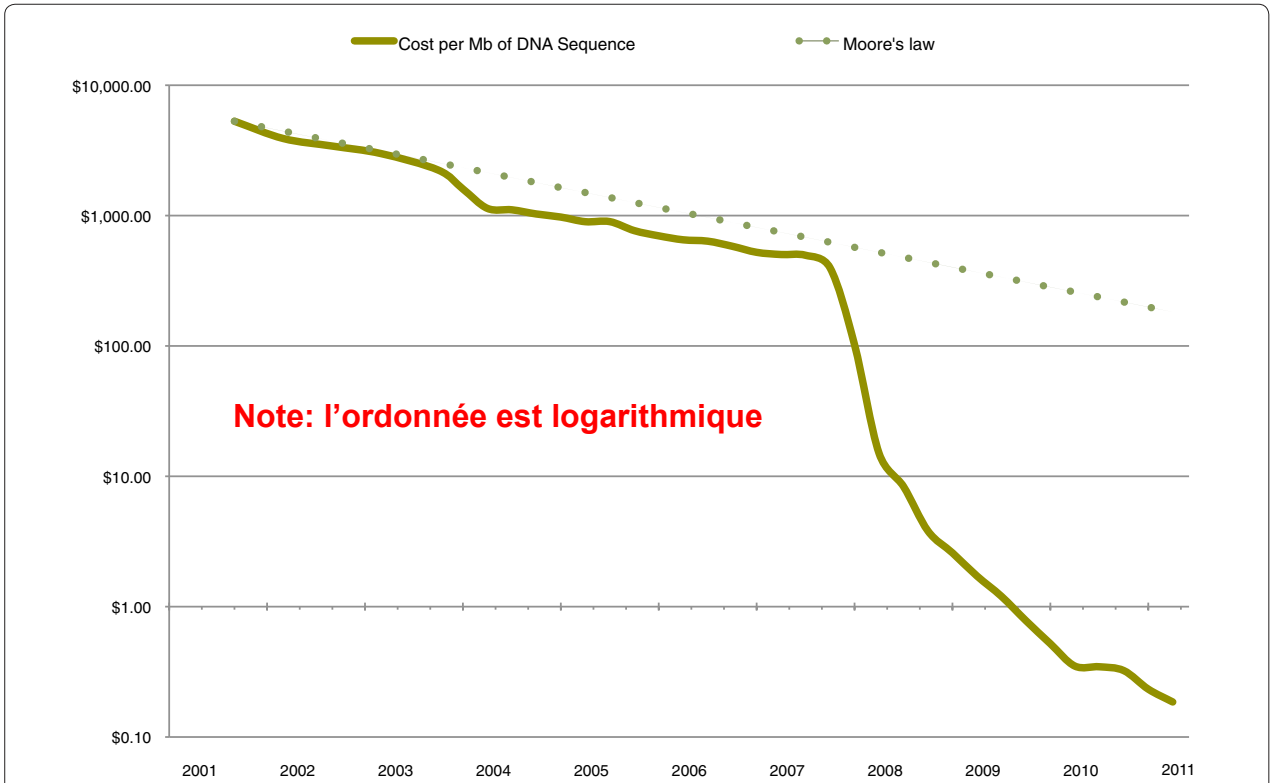


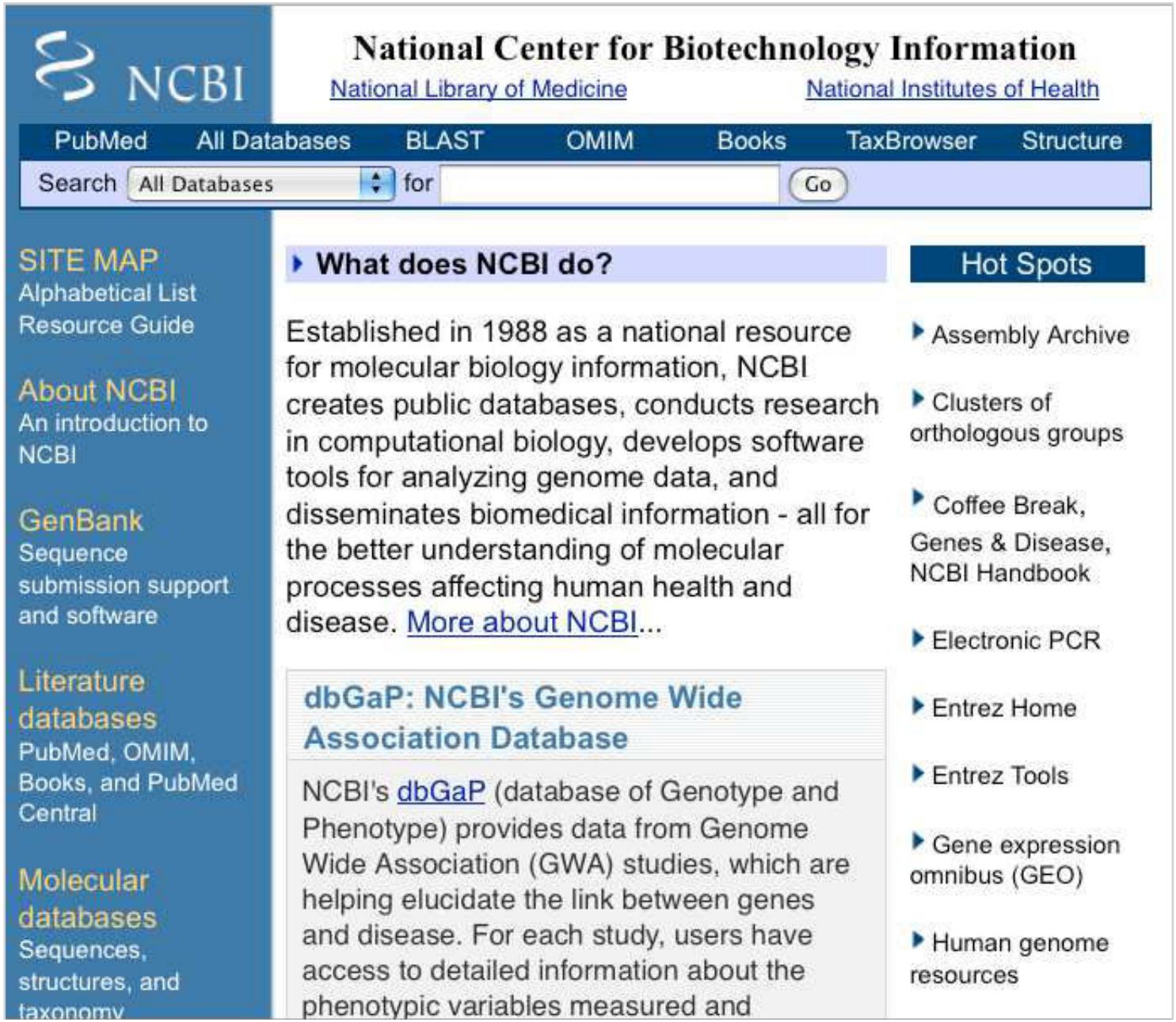
Figure 2. Cost of 1 MB of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore's law. Adapted from [11]. Cost was calculated in January of each year. MB, megabyte.

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

- Figure repdroduite d'après Sboner et al. The real cost of sequencing: higher than you think!. Genome Biol (2011) vol. 12 (8) pp. 125

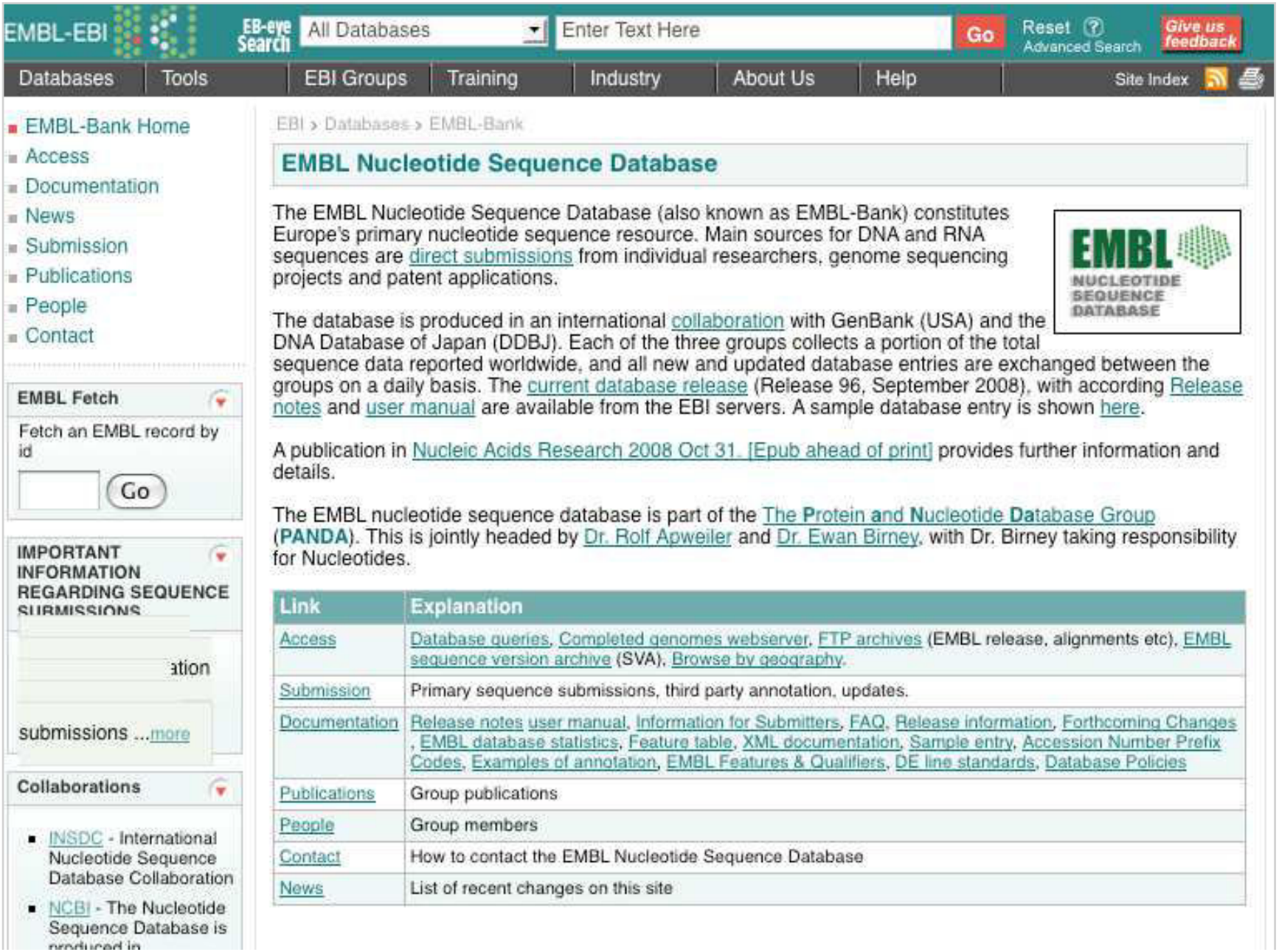
Genbank (NCBI - USA)

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



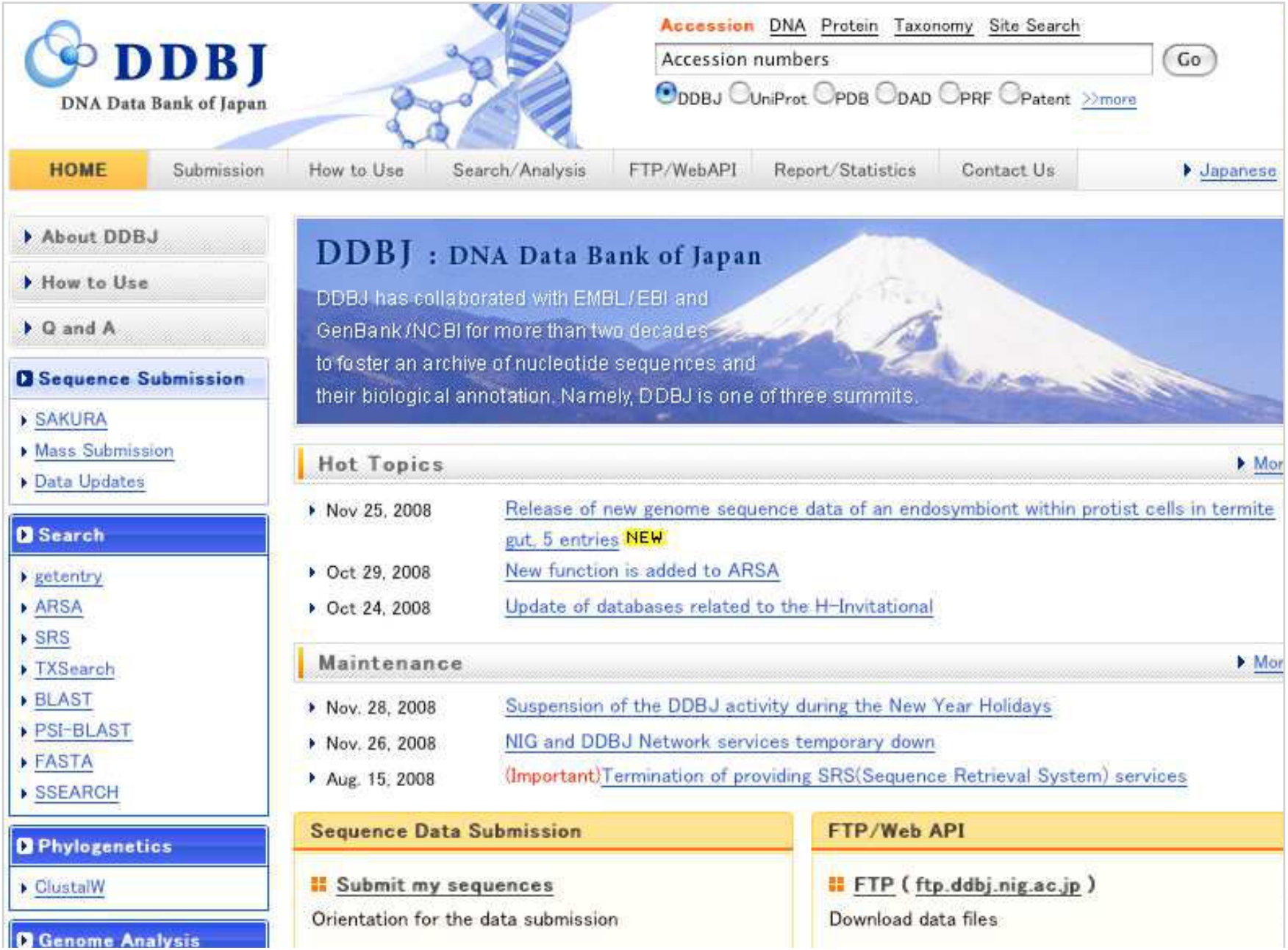
The EMBL Nucleotide Sequence Database (EBI - UK)

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



DDBJ - DNA Data Bank of Japan

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



Bases de données de séquences protéiques

Banques généralistes de séquences protéiques

- PIR (Protein Information Ressource)
  - Première banque des protéines (1965)
  - Banque américaine (NBRF- National Biomedical Research Fondation)
  - 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 2914, 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.

Uniprot

Swiss-Prot + PIR + TrEMBL-EBI

↓

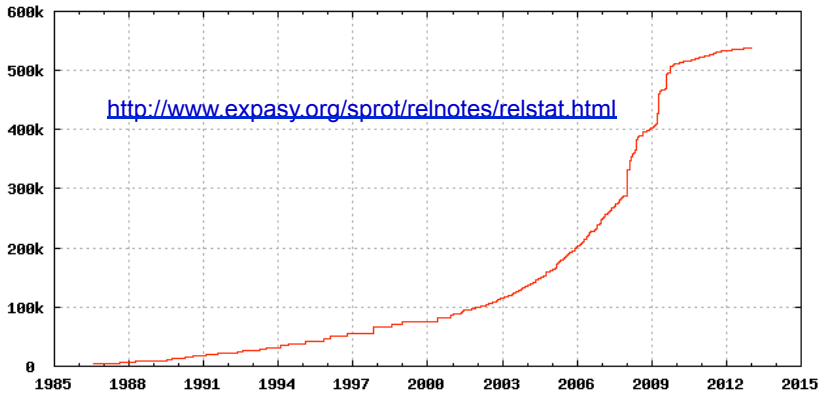
UniProt
(Universal Protein Ressource)
http://www.uniprot.org/

UniProt - the Universal Protein Resource

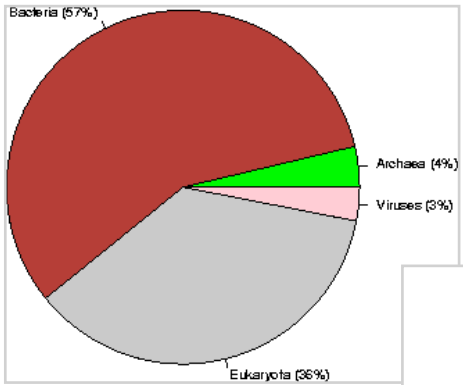
http://www.uniprot.org/

- 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.

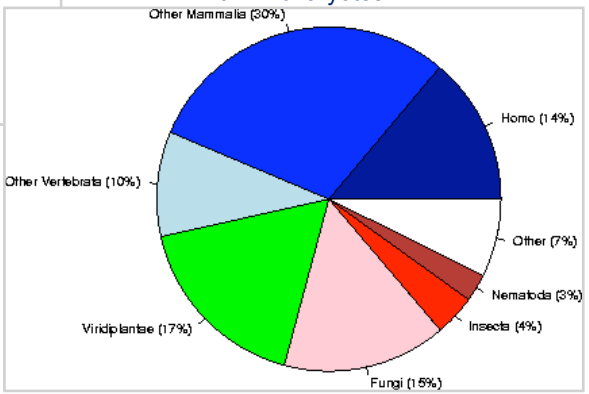
Number of entries (polypeptides) in **Swiss-Prot**



Taxonomic distribution of the sequences



Within Eukaryotes



UniProt example - Human Pax-6 protein
Header : name and synonyms

Reviewed, UniProtKB/Swiss-Prot **P26367** (PAX6_HUMAN)

Last modified November 25, 2008. Version 110. History...

Contribute
Send feedback

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:* **Paired box protein Pax-6** *Alternative name(s):* 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.

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. 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. 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. 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. 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. 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. Defects in PAX6 are a cause of coloboma of optic nerve [MIM:120430]. 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. Contains 1 homeobox DNA-binding domain. Contains 1 paired domain.

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	cell differentiation Inferred from electronic annotation. Source: UniProtKB-KW central nervous system development Traceable author statement. Source: ProtInc eye development Traceable author statement. Source: ProtInc organ morphogenesis Ref.19 Traceable author statement. Source: ProtInc regulation of transcription, DNA-dependent Inferred from electronic annotation. Source: InterPro visual perception Ref.19 Traceable author statement. Source: ProtInc
Cellular component	nucleus Inferred from electronic annotation. Source: InterPro
Molecular function	protein binding Inferred from physical interaction. Source: IntAct sequence-specific DNA binding Inferred from electronic annotation. Source: InterPro transcription factor activity Ref.19 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. 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	ENSG00000007372. Homo sapiens. [Contig view]
GeneID	5080.
KEGG	hsa:5080.

Structure tridimensionnelle des protéines

PDB - The Protein Data Bank

http://www.rcsb.org/pdb/

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Nov 25, 2008 there are 54466 Structures | PDB Statistics

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.

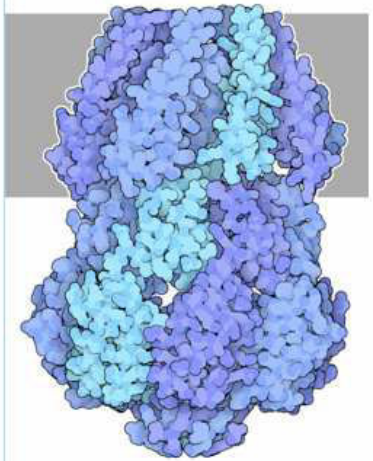
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

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

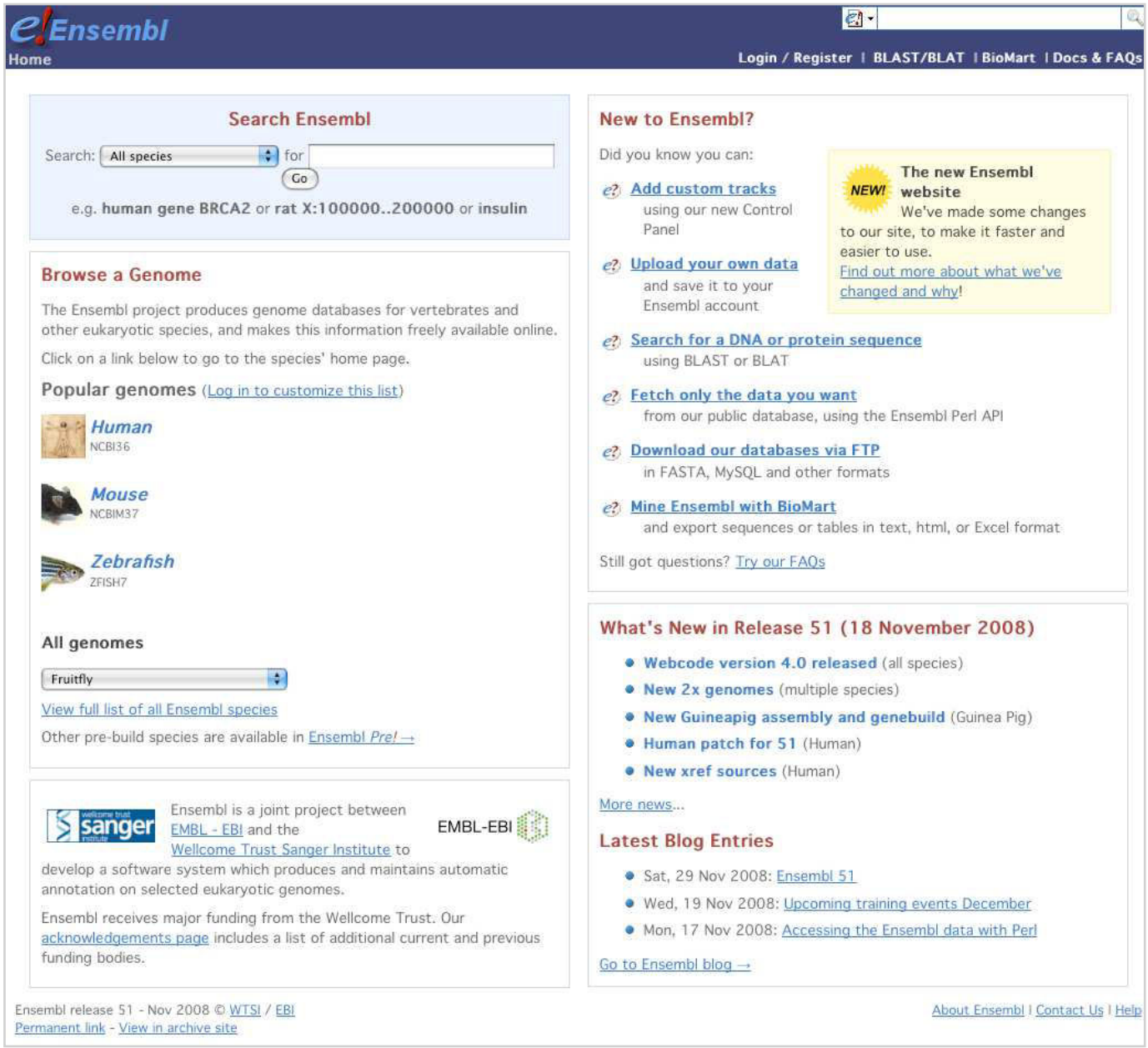
During the past year, the wwPDB annotators have collaborated on a project to clarify the details and procedures related to data processing and annotation.

- Full article ...

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

EnsEMBL Genome Browser (Sanger Institute + EBI)

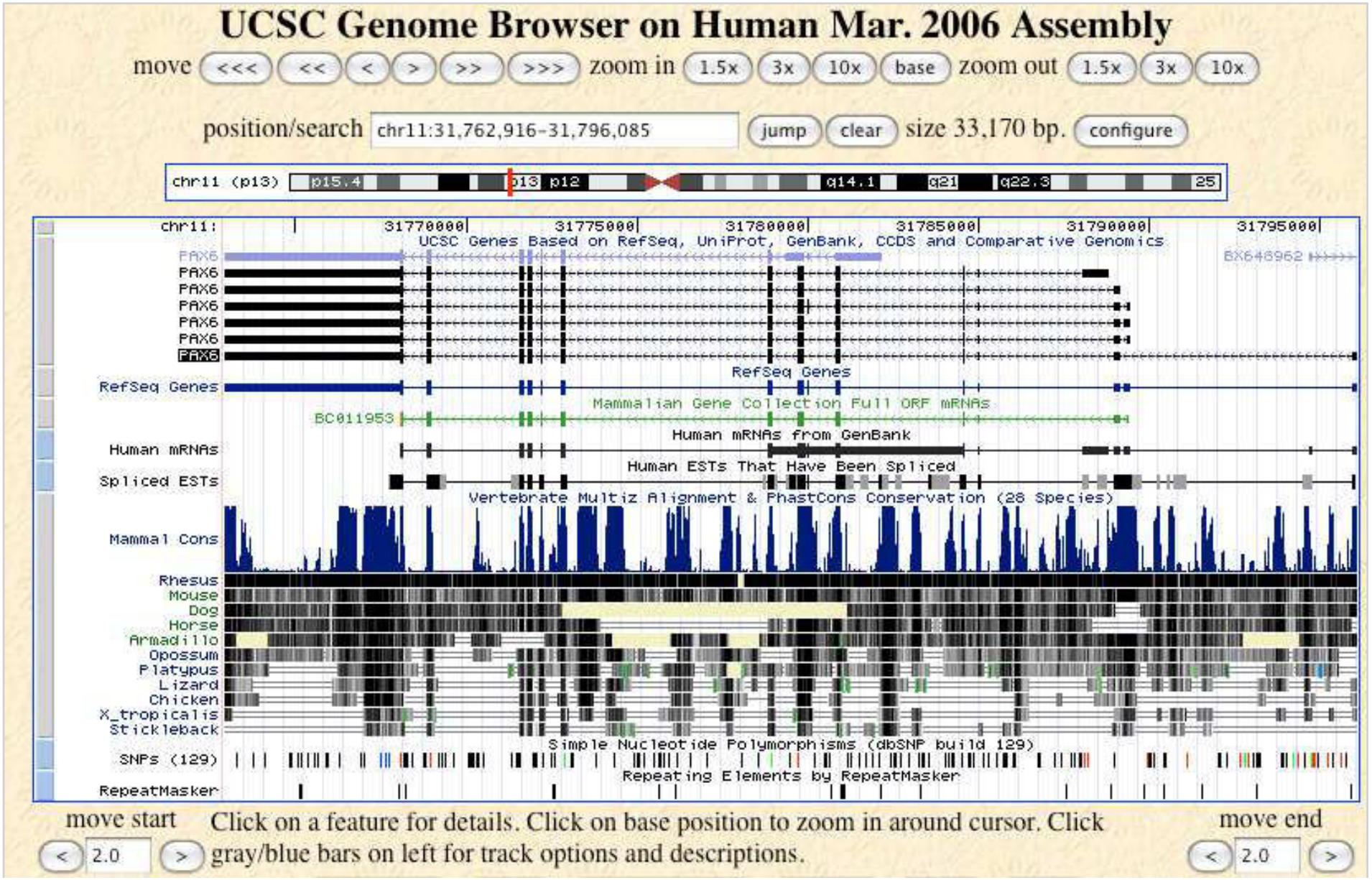
http://www.ensembl.org/



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

http://genome.ucsc.edu/

Human gene Pax6 aligned with Vertebrate genomes

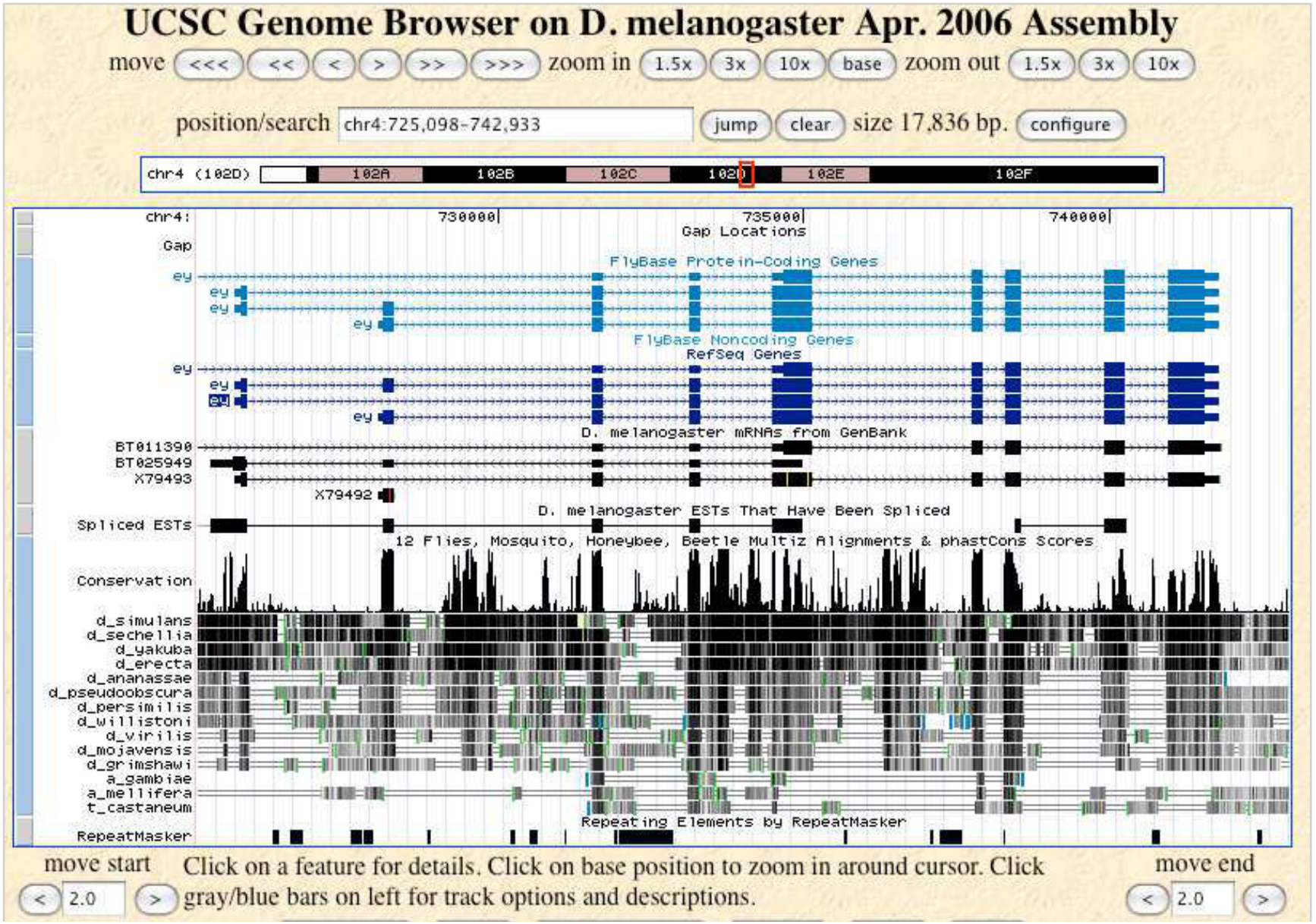


UCSC Genome Browser *(University California Santa Cruz - USA)*

http://genome.ucsc.edu/

UCSC

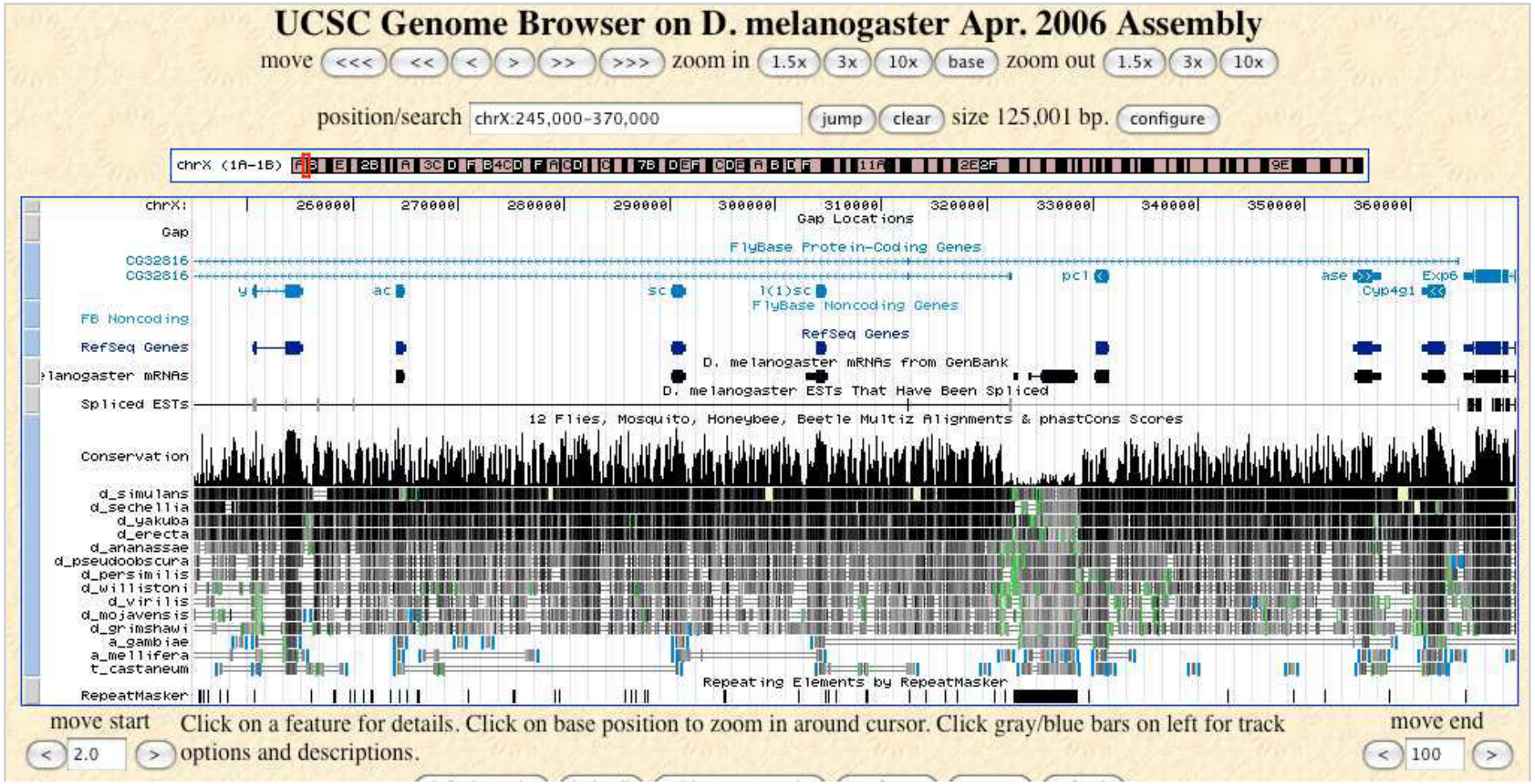
Drosophila gene eyeless (homolog to Pax6) aligned with Insect genomes



UCSC Genome Browser (University California Santa Cruz - USA)

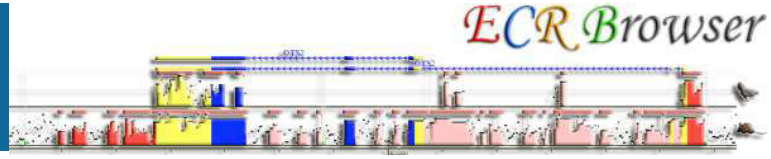
http://genome.ucsc.edu/

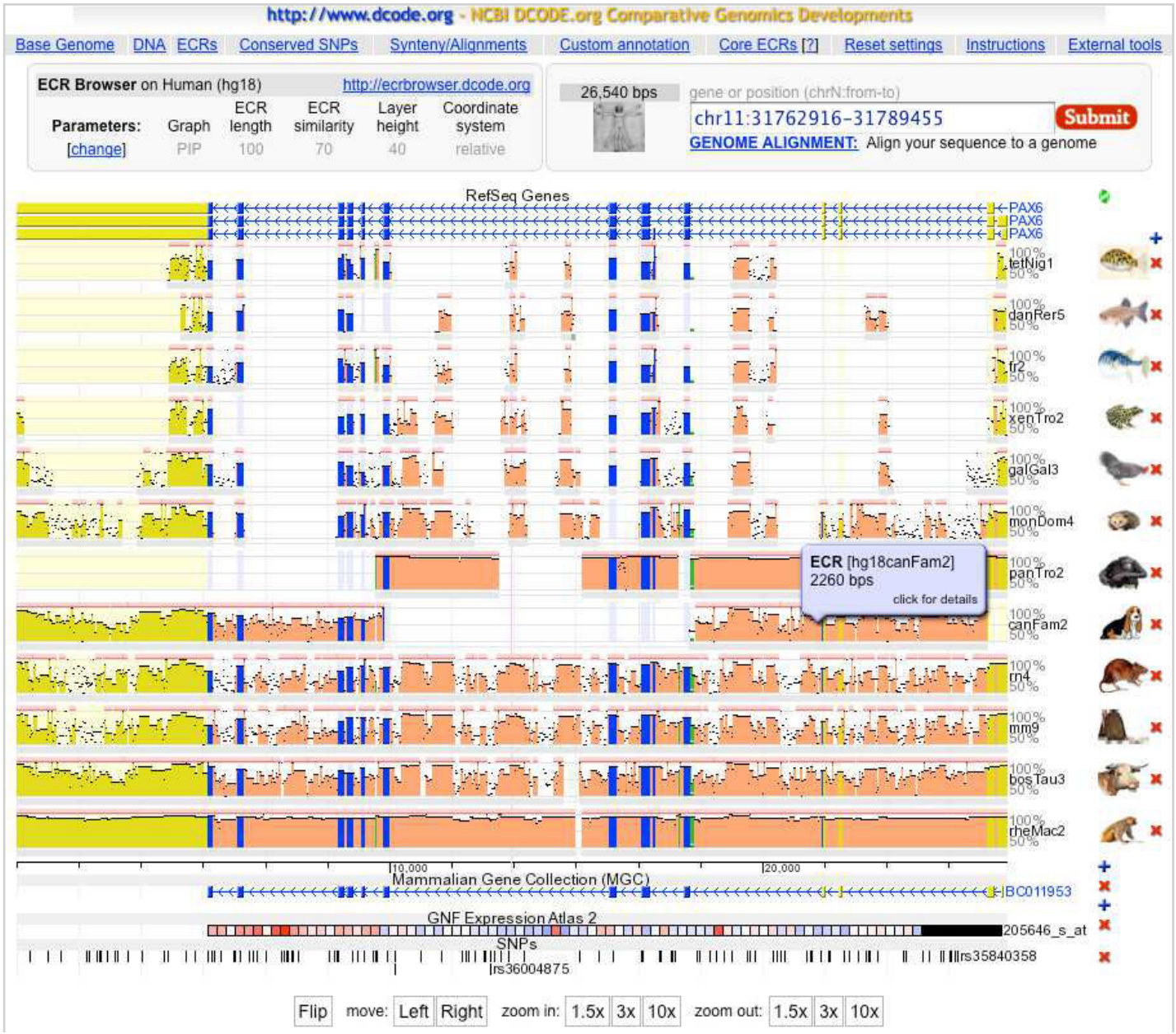
Drosophila 120kb chromosomal region covering the Achaete-Scute Complex



ECR Browser

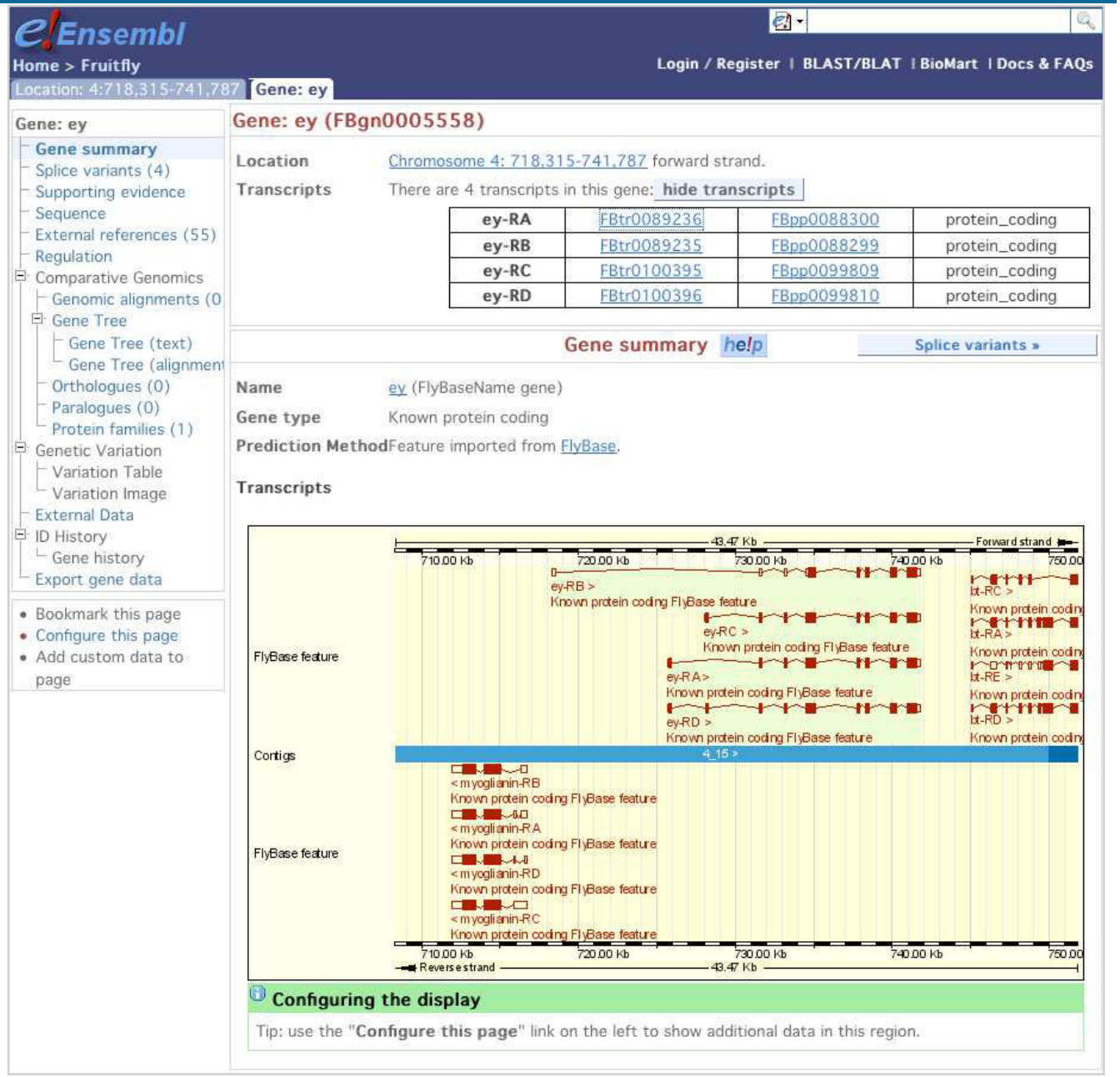
http://ecrbrowser.dcode.org/





EnsEMBL - Example: Drosophila gene Pax6

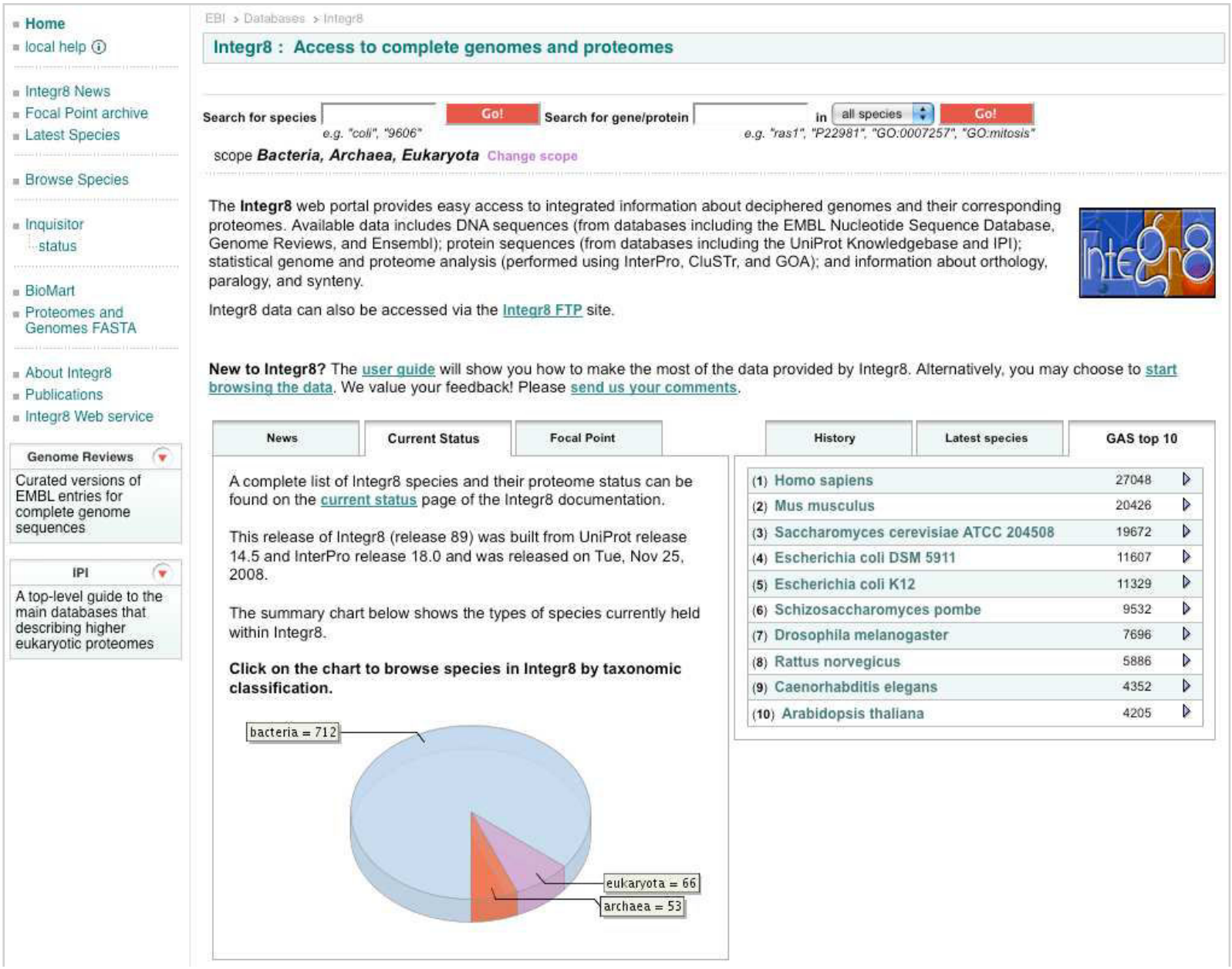
http://www.ensembl.org/



Génomique comparative

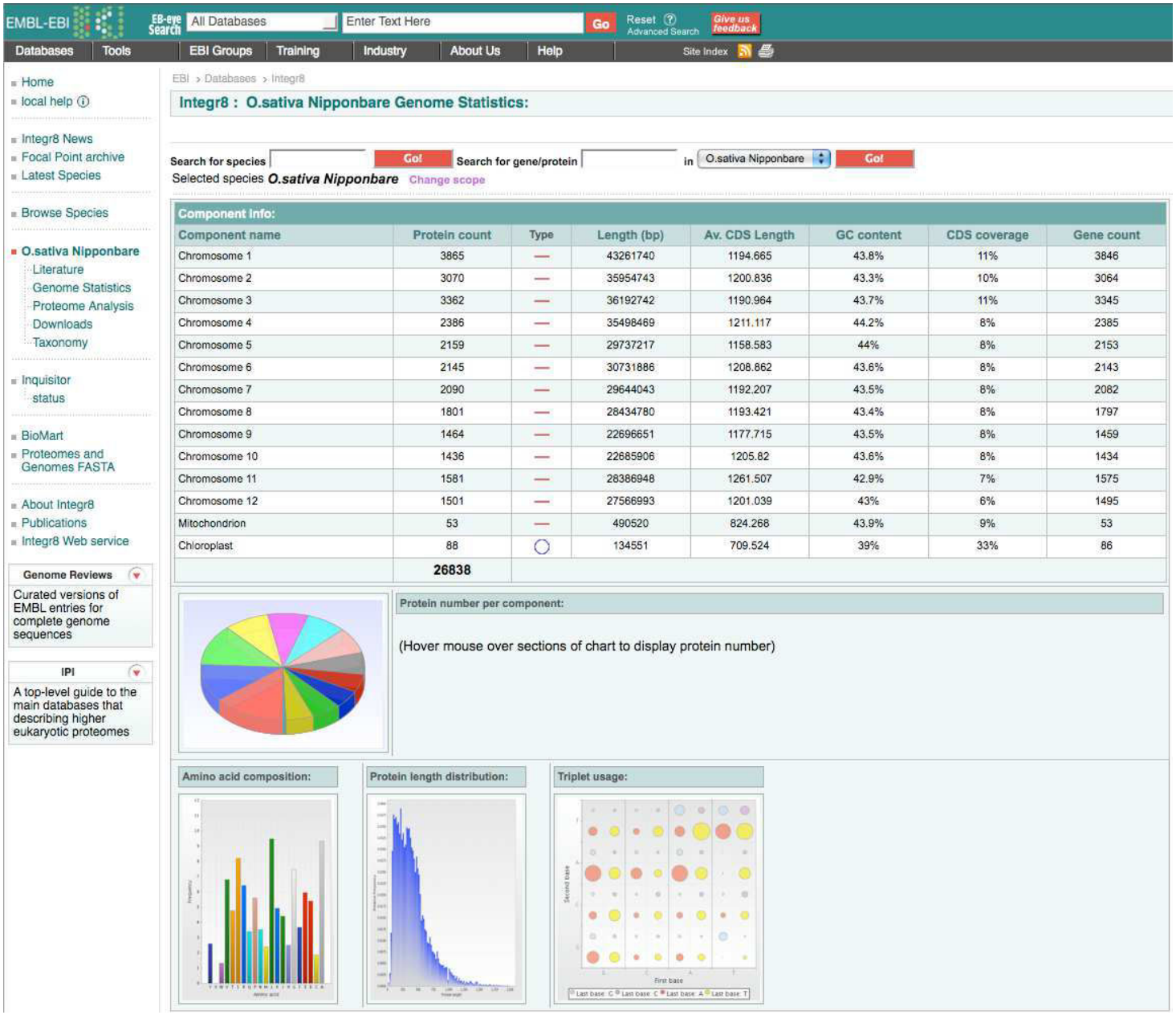
Integr8 - access to complete genomes and proteomes

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



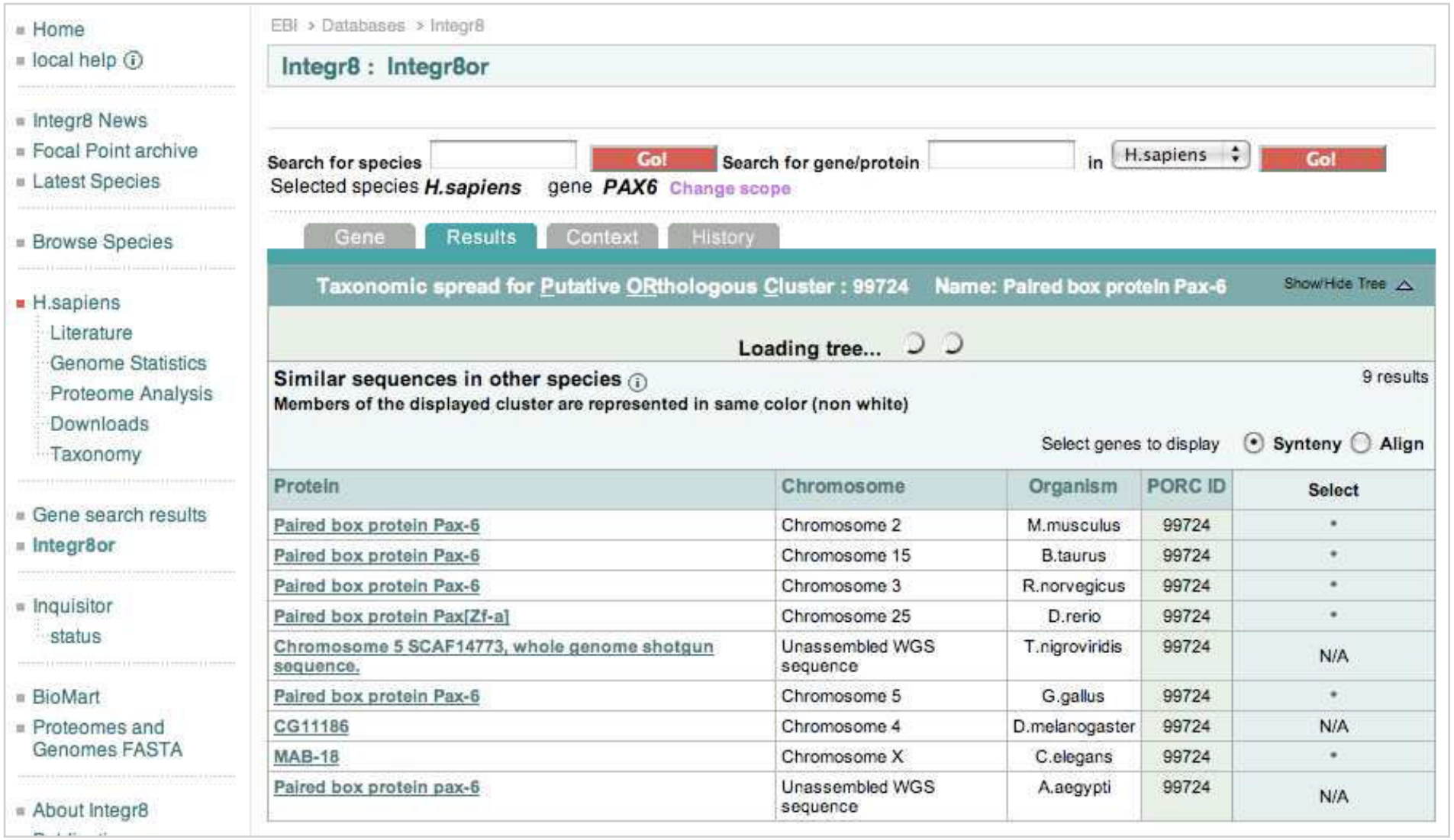
Integr8 - genome summaries

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



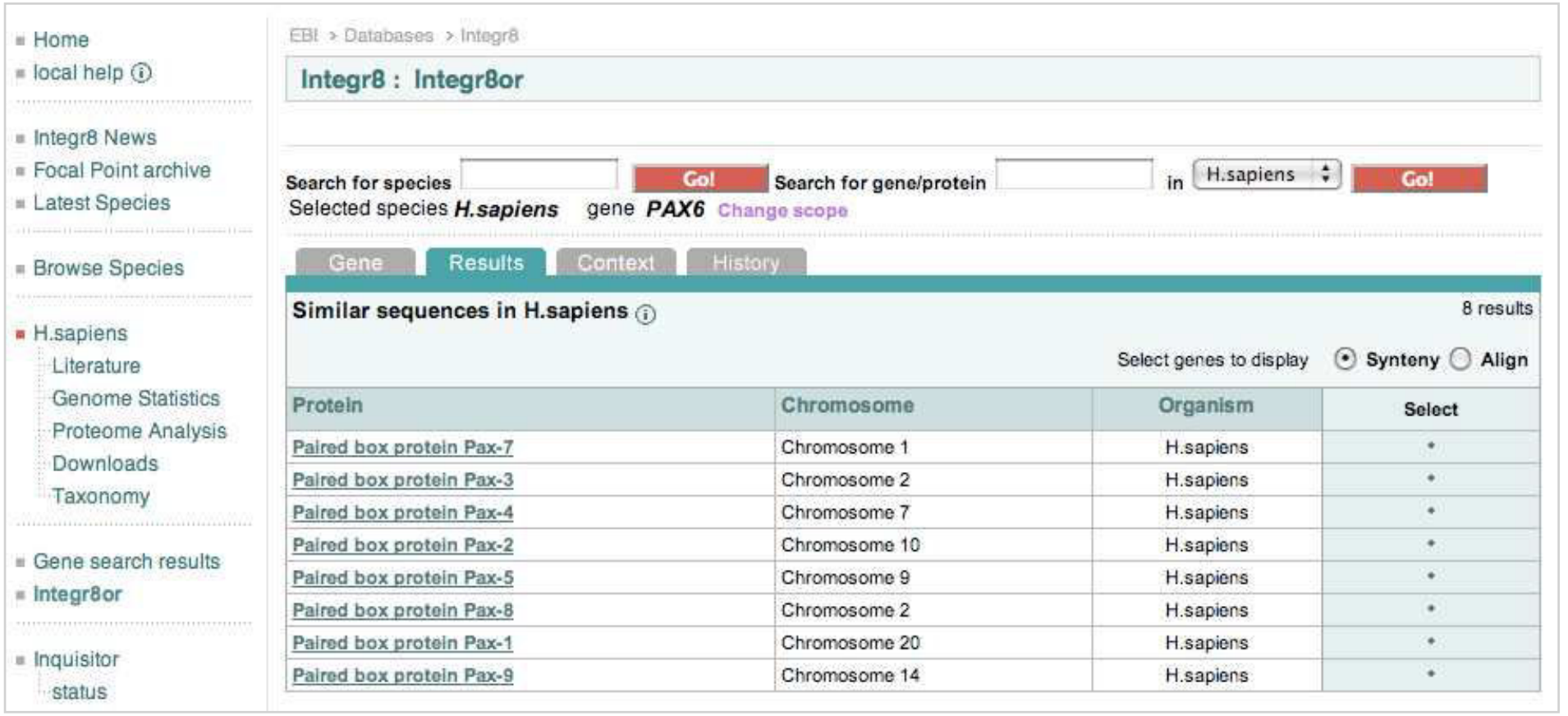
Integr8 - clusters of orthologous genes (COGs)

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



Integr8 - clusters of paralogous genes

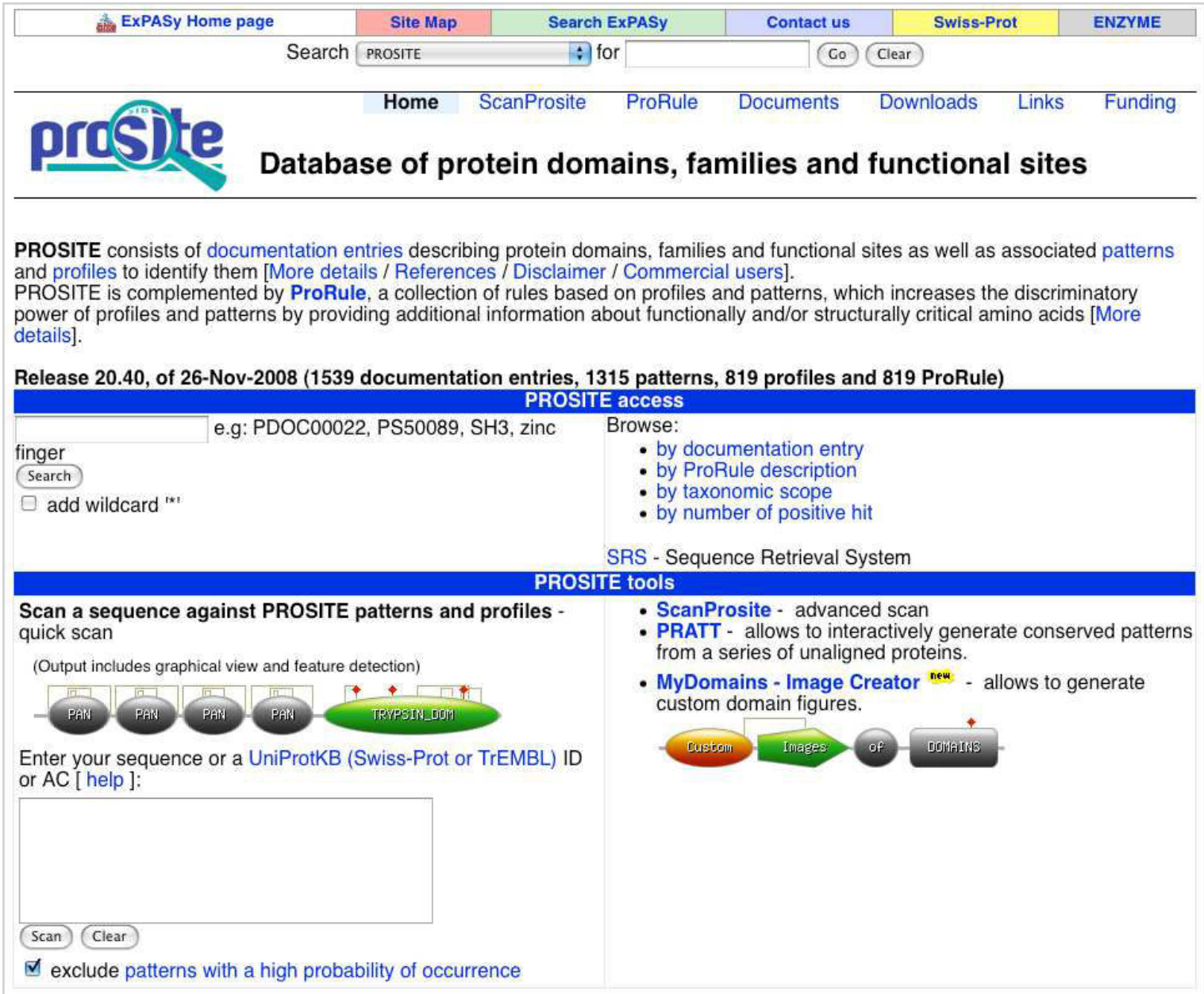
http://www.ebi.ac.uk/integr8/



Domaines protéiques

Prosite - protein domains, families and functional sites

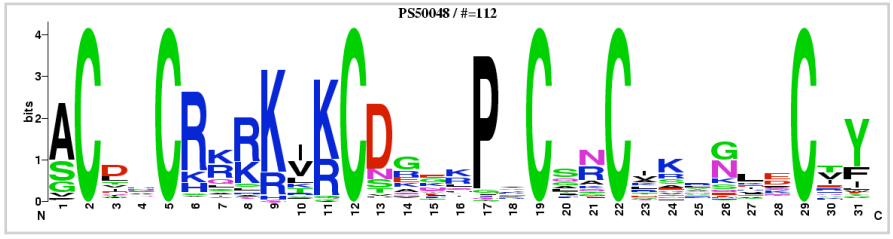
http://www.expasy.ch/prosite/



Prosite – séquences alignées et logos

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

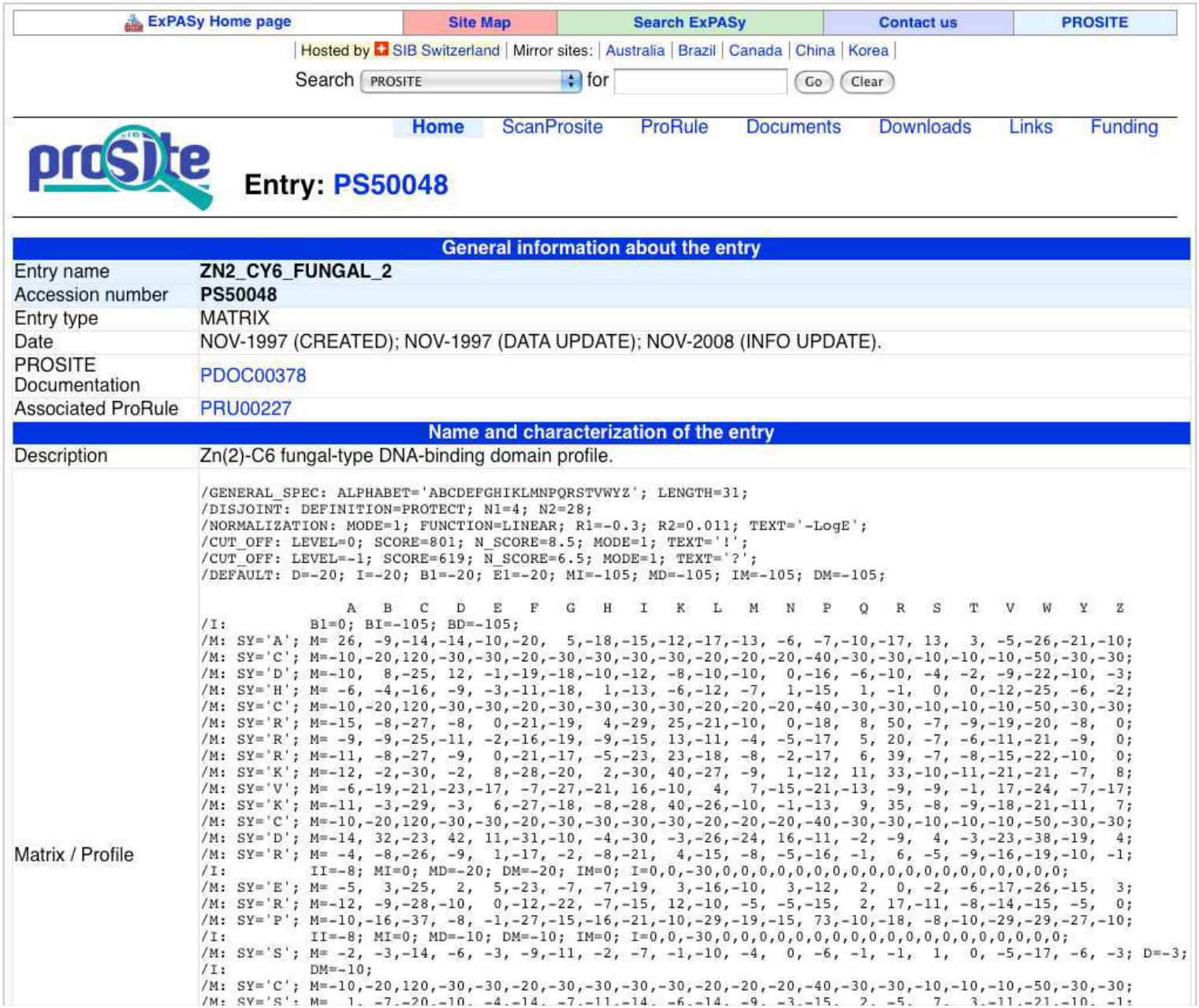
- La figure ci-dessous représente un profil de domaine conservé (graphique dit « ***logo*** »). Il indique la conservation des acides aminés à chaque position sein d'un domaine protéique.
- Le logo est obtenu à partir de l'alignement d'une série de séquences protéiques.
- Illustration
  - Profil "Prosite" 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_TRIRE/6-38          ACDRCHDKKLRCPRi.sGSP........CCSRCAKANV...ACVF
ACR2_NEUCR/21-49         ACYNCHRKRLRCDK...SLP........ACLKCSINGE...EC--
ACU15_NEUCR/23-53        ACDRCRSKKIRCDG...IRP........CCSQCANVGF...ECKT
AFLR_ASPFL/28-58         SCTSCASSKVRCTK...EKP........ACARCIERGL...ACQY
AFLR_ASPPA/28-58         SCTSCASSKVRCTK...EKP........ACARCIERGL...ACQY
AFLR_EMENI/27-57         SCISCSRSKVKCNK...EKP........TCSRCVRRGL...PCEY
ALCR_EMENI/11-51         SCDPCRKGKRRCDAp.eNRNeanengwvSCSNCKRWNK...DCTF
AMDR_ASPFU/26-59         ACVHCHRRKVRCDArivGLP........-CSNCRSSGKt..DCRI
AMDR_ASPOR/25-58         ACIHCHRRKVRCDArivGLP........-CSNCRSAGKa..DCRI
AMDR_EMENI/19-52         ACVHCHRRKVRCDArlvGLP........-CSNCRSAGKt..DCQI
ARGR2_YEAST/20-48        GCWTCRGRKVKCDL...RHP........HCQRCEKSNL...PC--
ARO80_YEAST/24-60        ACISCRSRKVKCDLgpvDNphd.....pPCARCKRELK...KCIF
ATG2_PICPA/631-660       GCLTCRKRQVKCDE...RKP........FCLNCEKSEQ...KCT-
CAT8_YEAST/69-99         ACDRCRSKKTRCDG...KRP........QCSQCAAVGF...ECRI
CBF3B_YEAST/13-44        PCSVCTRRKVKCDR...MIP........-CGNCRKRGQd.sECMK
CHA4_YEAST/43-72         ACQNCRRRRRKCNM...EKP........-CSNCIKFRT...ECVF
CTF1A_FUSSO/60-92        ACETCHARKVRCDAaslGVP........-CTNCVAFQI...ECRI
CTF1B_FUSSO/52-83        ACVSCRARKVRCDVv.eGAP........-CGNCRWDNV...ECVV
CZF1_CANAL/317-347       GCLTCRQRKKRCCE...TRP........RCTECTRLRL...NCTW
DAL81_YEAST/149-181      SCNQCRLKKTKCNYf.pDLG........NCLECETSRT...KCTF
ECM22_YEAST/43-73        GCDNCKRRRVKCDE...GKP........FCKKCTNMKL...DCVY
FCR1_CANAL/25-54         ACDSCRIKKTKCDG...KKP........-CNRCTLDNK...ICVF
FLUF_NEUCR/10-39         ACLVCRKKRTKCDG...QMP........-CRRCRSRGE...ECAY
GAL4_YEAST/10-40         ACDICRLKKLKCSK...EKP........KCAKCLKNNW...ECRY
GRT1_SCHPO/13-42         ACENCRKRKVKCSG...GDV........-CFECQKYNE...NCVY
HAL9_YEAST/135-168       ACDHCRKRKIRCDEv..DQQt......kKCSNCIKFQL...PCTF
HAP1_YEAST/63-95         SCTICRKRKVKCDK...LRP........HCQQCTKTGVa.hLCHY
LAC9_KLULA/94-124        ACDACRKKKWKCSK...TVP........TCTNCLKYNL...DCVY
LEUR_YEAST/36-69         ACVECRQQKSKCDAh.eRAPe.......PCTKCAKKNV...PCIL
LYS14_YEAST/158-188      GCSECKRRRMKCDE...TKP........TCWQCARLNR...QCVY
MAL13_YEAST/12-41        ACDCCRIRRVKCDG...KRP........-CSSCLQNSL...DCTY
MAL33_YEAST/7-36         ACDYCRVRRVKCDG...KKP........-CSRCIEHNF...DCTY
MAL63_YEAST/7-36         SCDCCRVRRVKCDR...NKP........-CNRCIQRNL...NCTY
MOC3_SCHPO/35-63         GCLTCRRRRIKCDE...TKP........FCLNCTKTNR...EC--
NIRA_EMENI/41-72         ACIACRRRKSKCDG...NLP........SCAACSSVYHt..TCVY
NIT4_NEUCR/52-83         ACIACRRRKSKCDG...ALP........SCAACASVYGt..ECIY
OAF1_YEAST/65-95         VCQACWKSKTKCDR...EKP........ECGRCVKHGL...KCVY
PDR1_YEAST/45-74         ACDNCRKRKIKCNG...KFP........-CASCEIYSC...ECTF
PDR3_YEAST/14-43         ACVNCRKRKIKCTG...KYP........-CTNCISYDC...TCVF
PDR8_YEAST/30-61         SCAFCRKRKLKCSQ...ARP........MCQQCVIRKLp..QCVY
PIP2_YEAST/24-54         VCQACRKAKTKCDQ...EKP........RCGRCTKQNL...FCIY
PPR1_YEAST/33-63         ACKRCRLKKIKCDQ...EFP........SCKRCAKLEV...PCVS
PRIB_LENED/19-52         ACTTCRAAKMKCVGa..EDGq......rQCQRCKRANV...QCIF
PRO1_NEUCR/54-84         GCYTCRLRRKKCDE...GSP........MCTACKHLGL...CCEY
PUT3_YEAST/33-62         ACLSCRKRHIKCPG...GNP........-CQKCVTSNA...ICEY
```

Prosite – Exemple de matrice de profil

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



ExPASy Home page | Site Map | Search ExPASy | Contact us | PROSITE

Hosted by SIB Switzerland | Mirror sites: | Australia | Brazil | Canada | China | Korea |

Search PROSITE for Go 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.

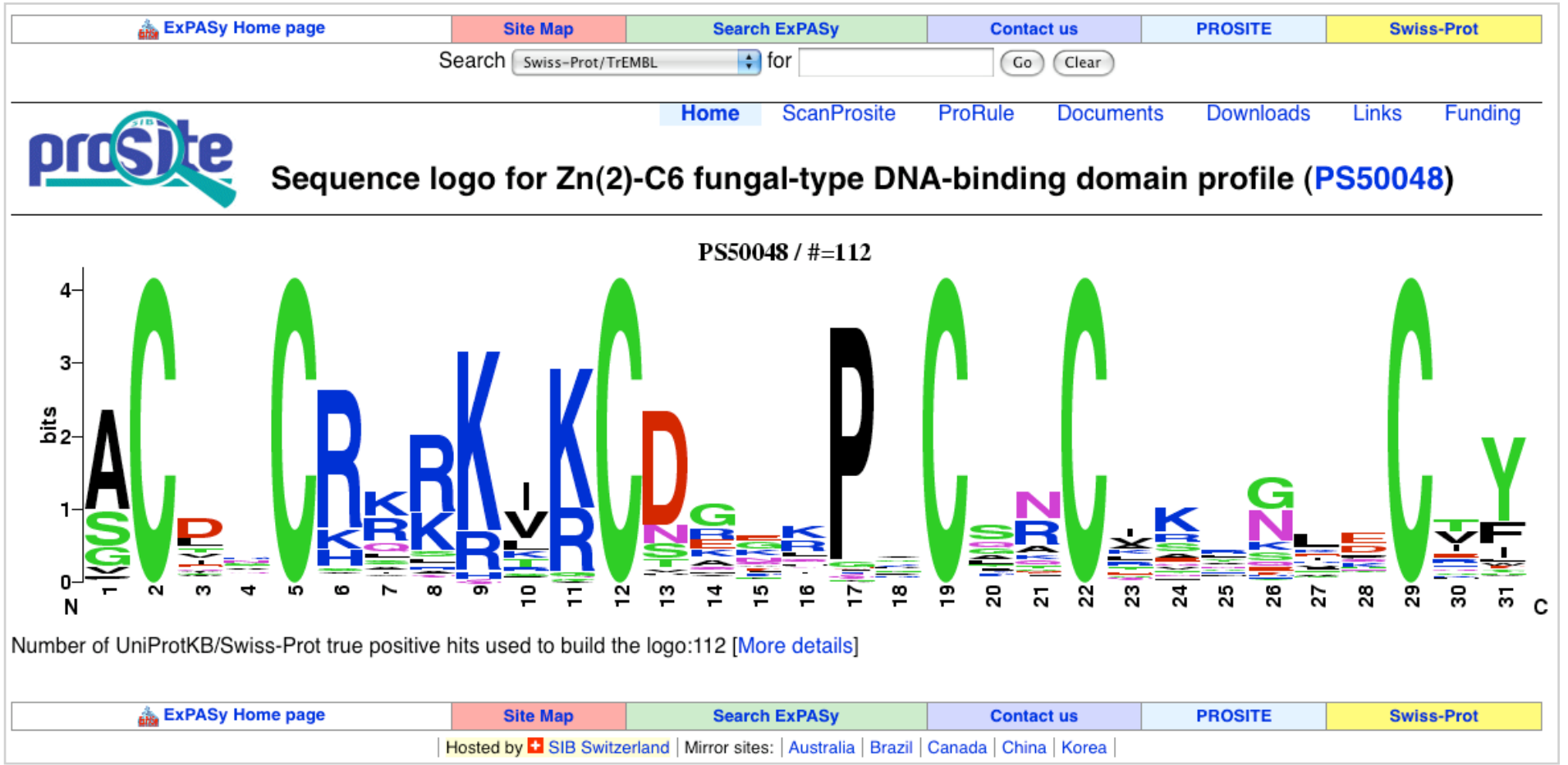
Matrix / Profile

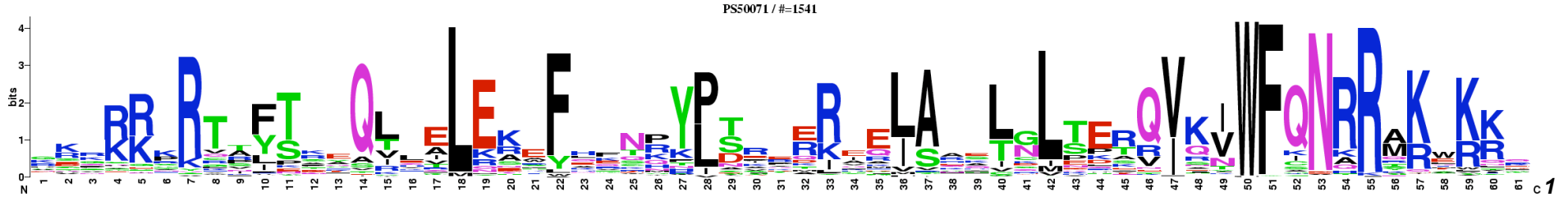
```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=31;
/DISJOINT: DEFINITION=PROTECT; N1=4; N2=28;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.3; R2=0.011; TEXT='-LogE';
/CUT_OFF: LEVEL=0; SCORE=801; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=619; N_SCORE=6.5; MODE=1; TEXT='?';
/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; BI=-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;
```

Prosite – Exemple de logo

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





Prosite - Example of domain signature

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

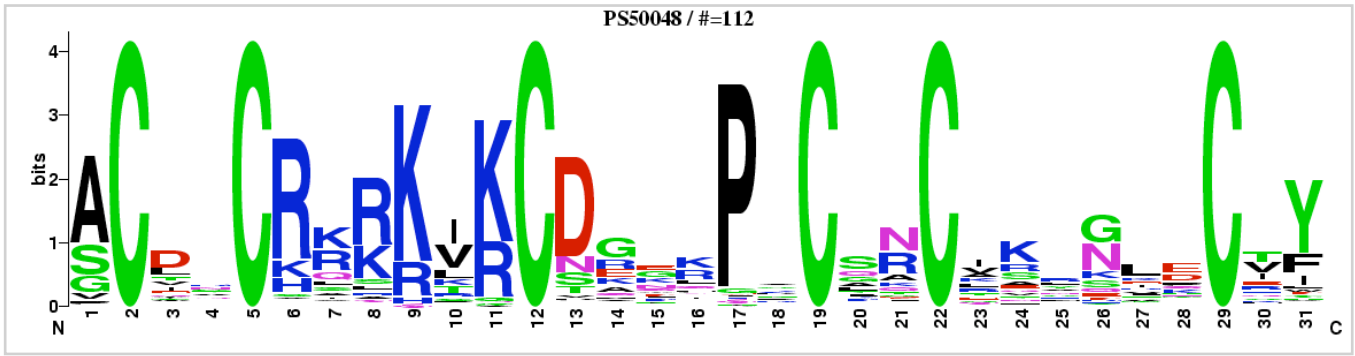
- The domain signature is a string-based pattern representing the residues that are characteristic of a domain.

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.

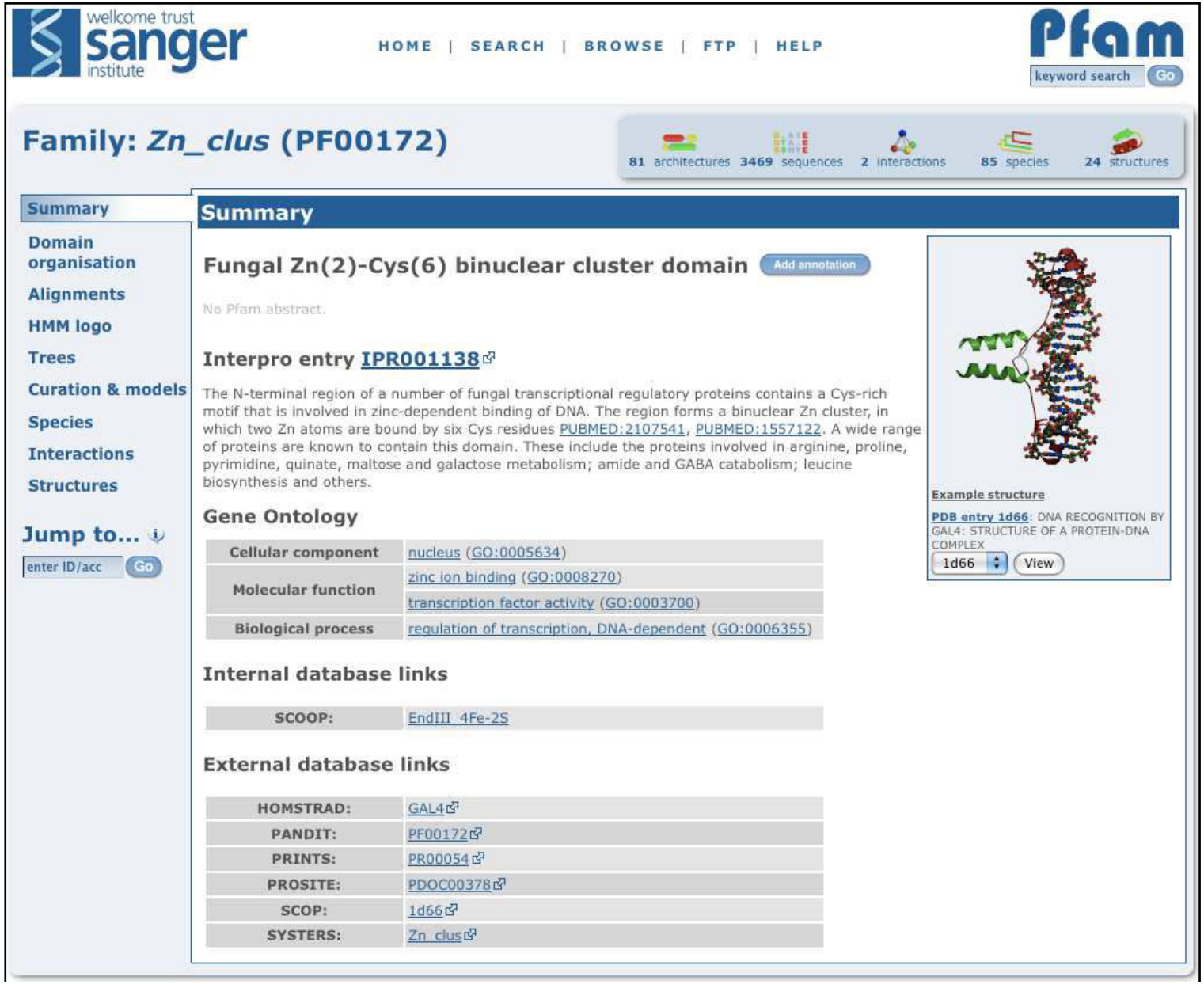
- Retrieve an alignment of Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS00463
- Retrieve a list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

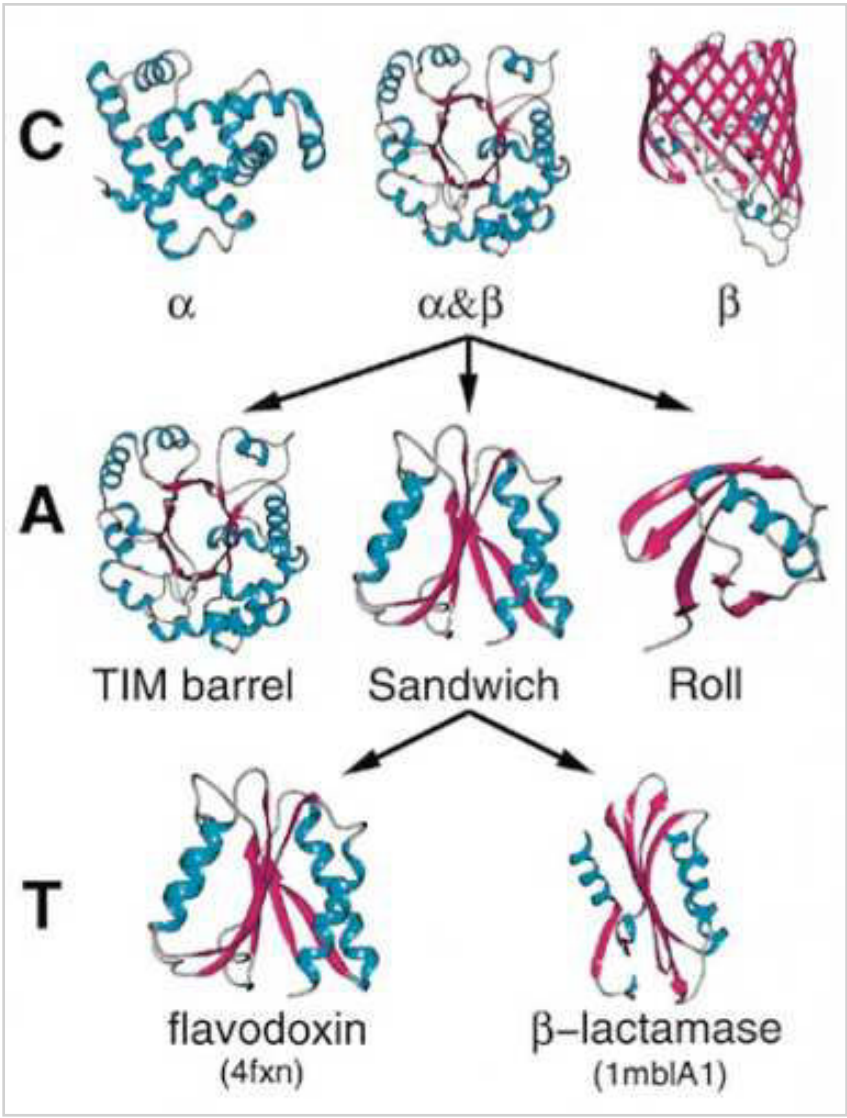
Protein families represented by multiple sequence alignments and hidden Markov models (HMMs)



CATH – Classification de structures protéiques

http://www.cathdb.info/

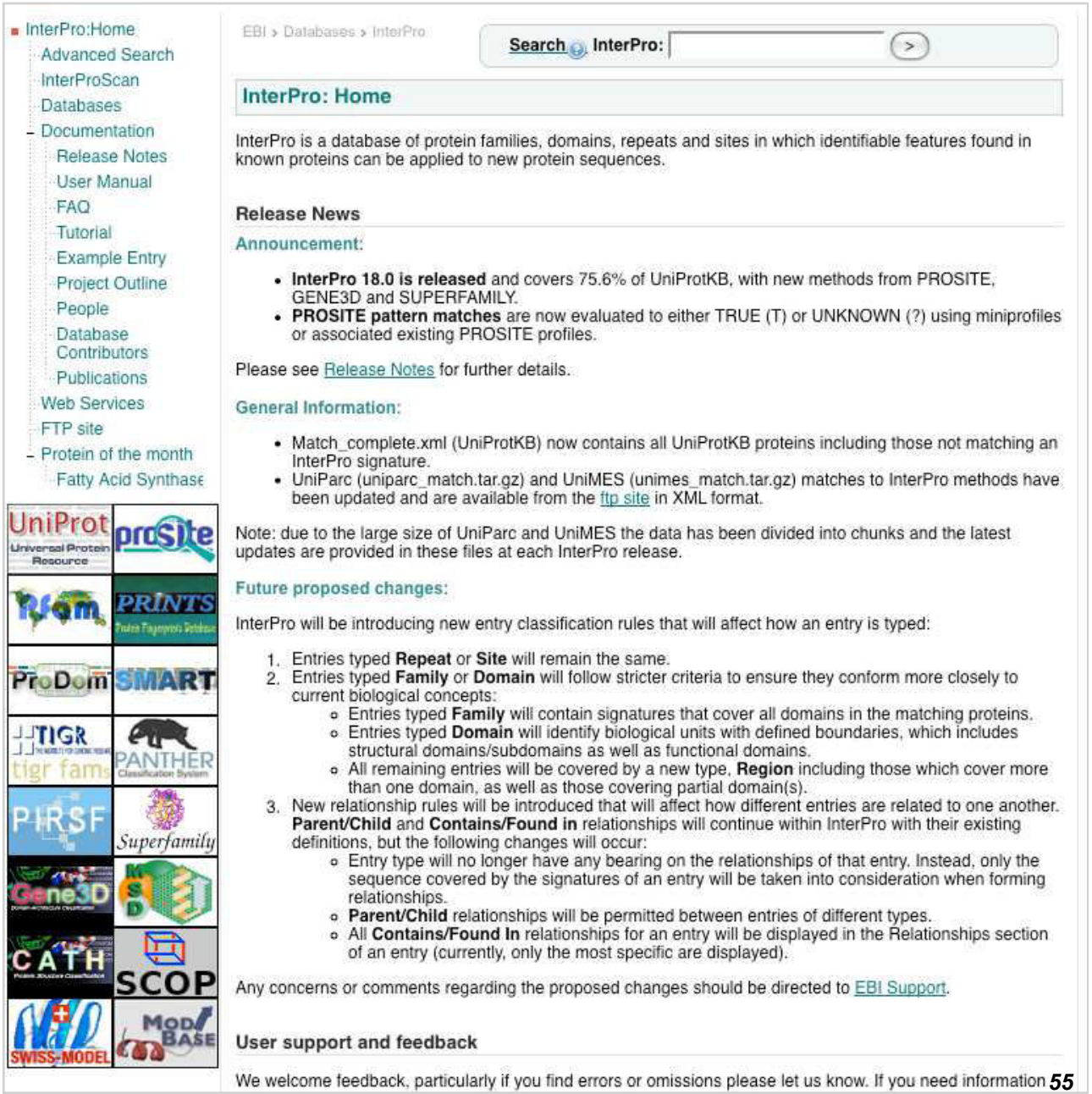
- CATH est une classification hiérarchique des domaines observés dans les structures tri-dimensionnelles de protéines.
- Les domaines sont classés selon 4 niveaux hiérarchiques
  - Classe (C),
  - Architecture (A),
  - Topologie (T)
  - Homologie (H).
- Les limites des classes et les assignations des 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.

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

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

EBI > Databases > InterPro

Search InterPro: >

Jump to: InterProScan Databases Documentation FTP site Help Advanced search

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

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

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

In vertebrates the corresponding Hox genes are known [4] as Hox-A2, A3, A4, A5, A6, A7, Hox-B1, B2, B3, B4, B5, B6, B7, B8, Hox-C4, C5, C6, C8, Hox-D1, D3, D4 and D8.

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

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

Structural links	PDB - click here SCOP: 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*

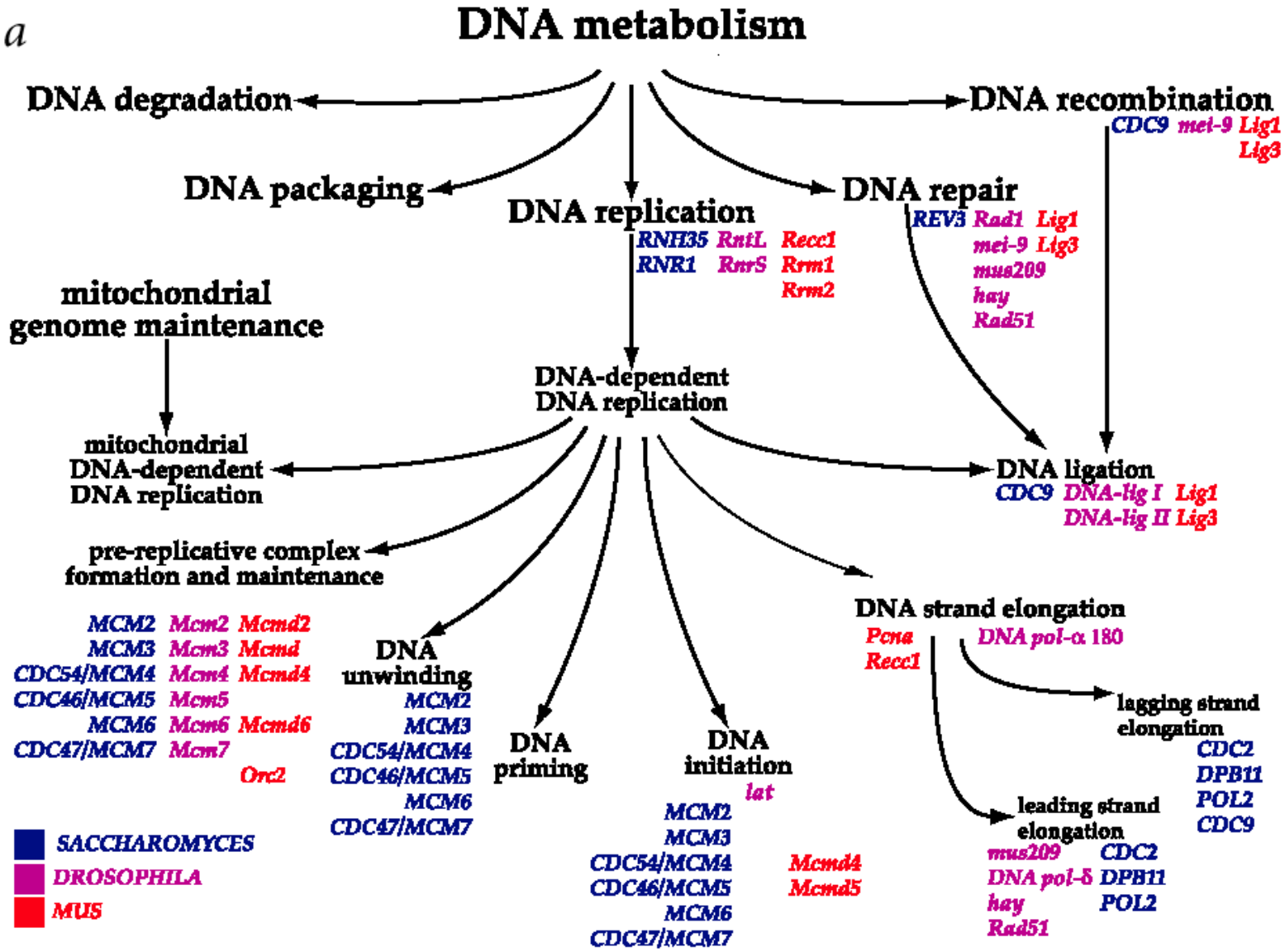
Les"bio-ontologies"

- 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.

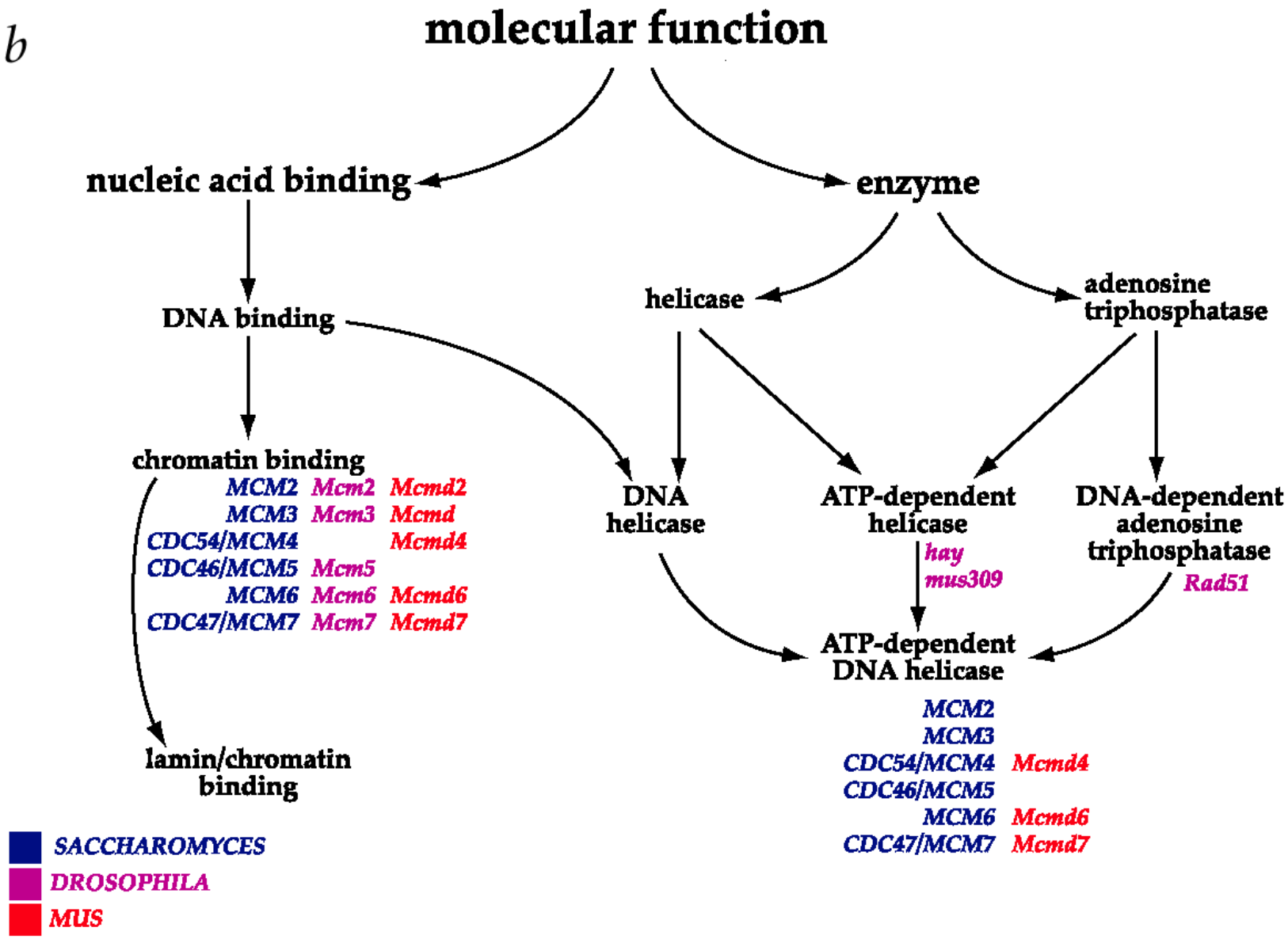
- Les bio-ontologies visent à répondre au problème d'inconsistance entre annotations.
- Pour y répondre, on met en place
  - Un vocabulaire contrôlé
    - On utilise toujours le même mot pour désigner le même concept.
    - Les listes de synonymes permettent d'établir les correspondances.
  - Classification 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, …).

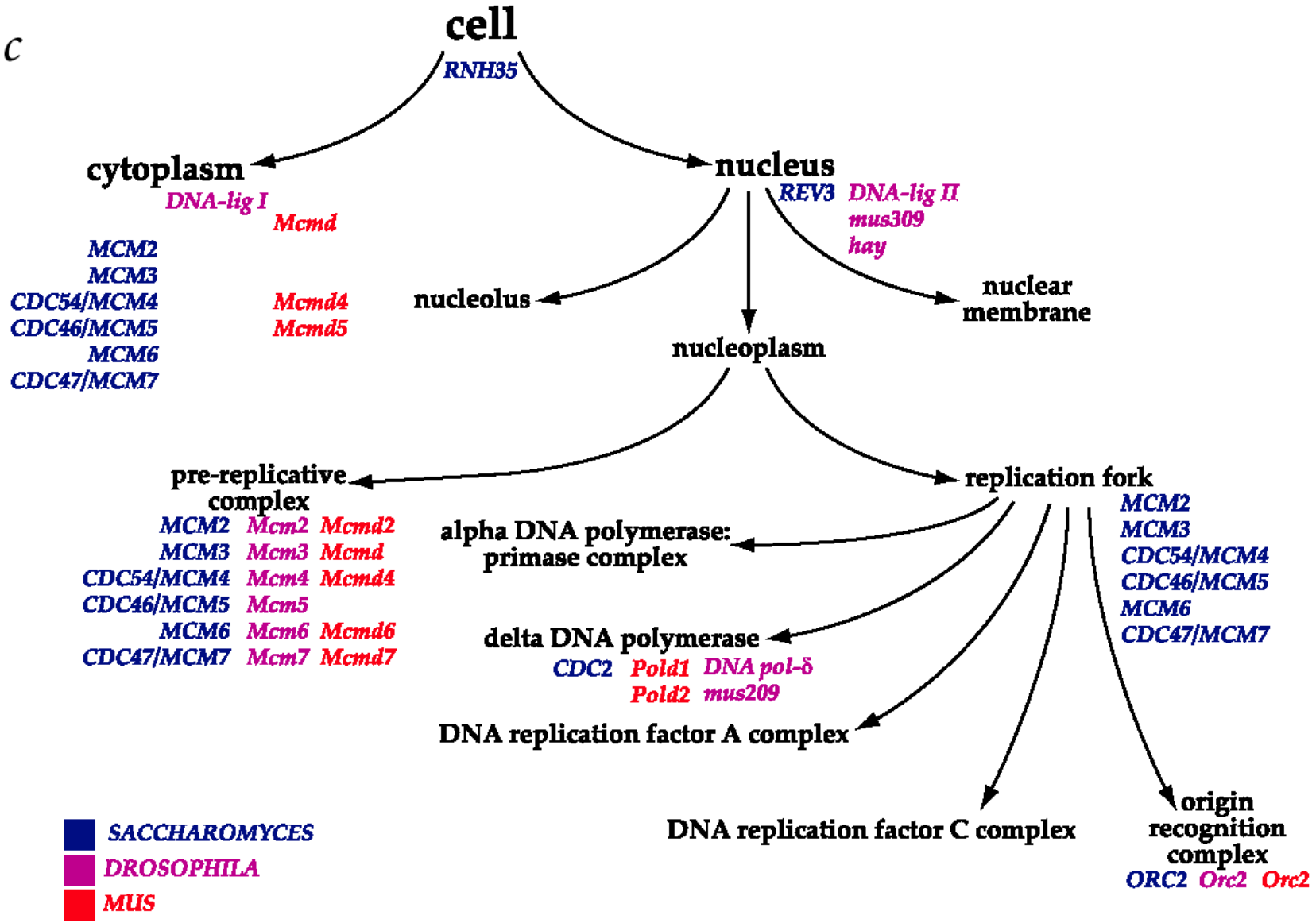
Gene ontology: exemples de processus biologiques



Gene ontology: exemples de fonctions moléculaires

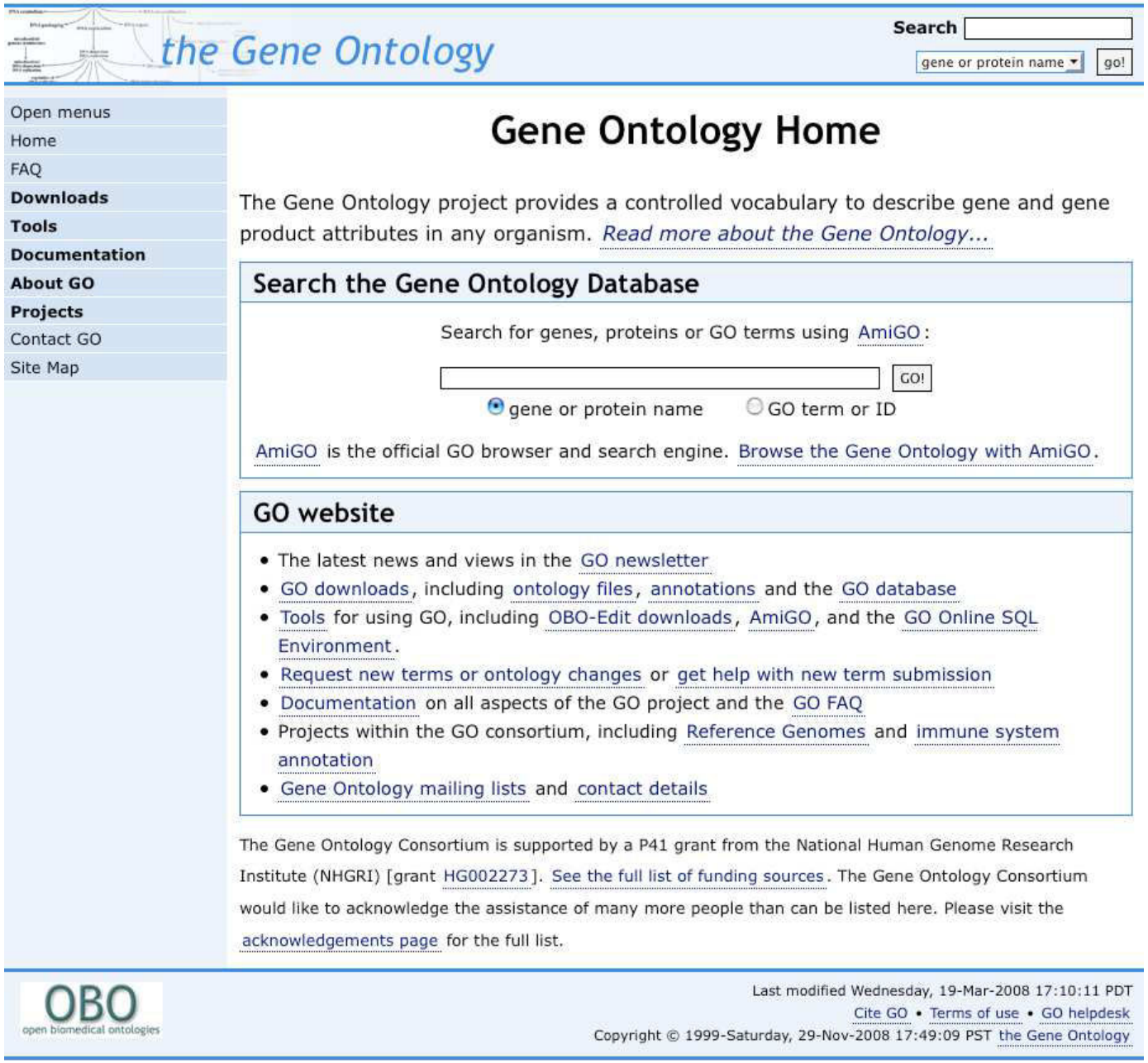


Gene ontology: exemples de composantes cellulaires



Gene Ontology Database

http://www.geneontology.org/



the Gene Ontology

Search

gene or protein name

go!

- Open menus
- Home
- FAQ
- **Downloads**
- **Tools**
- **Documentation**
- **About GO**
- **Projects**
- Contact GO
- Site Map

Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism. *Read more about the Gene Ontology...*

Search the Gene Ontology Database

Search for genes, proteins or GO terms using AmiGO:

GO!

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.

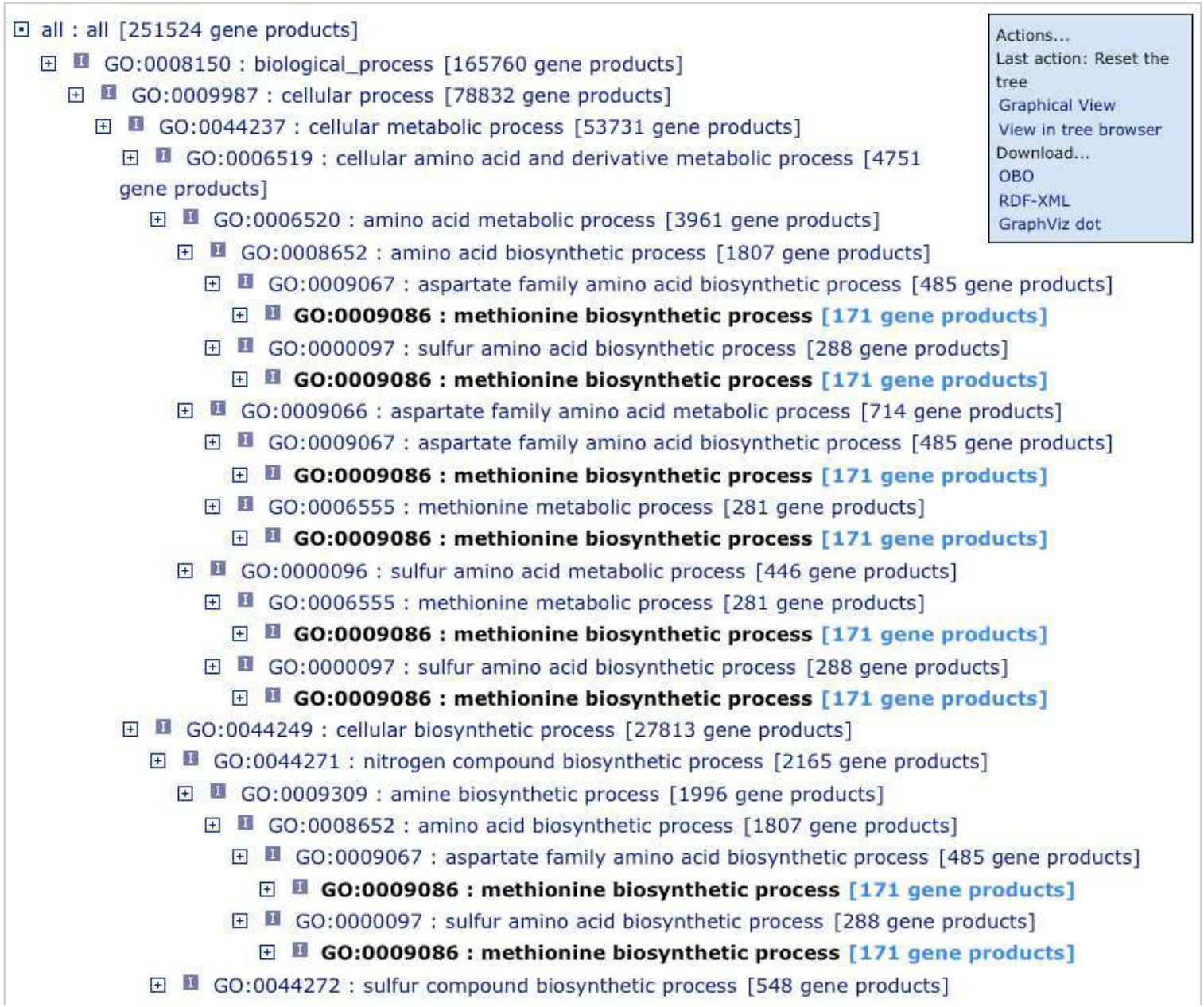
OBO open biomedical ontologies

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



Statut des annotations de GO (NAR DB issue 2006)

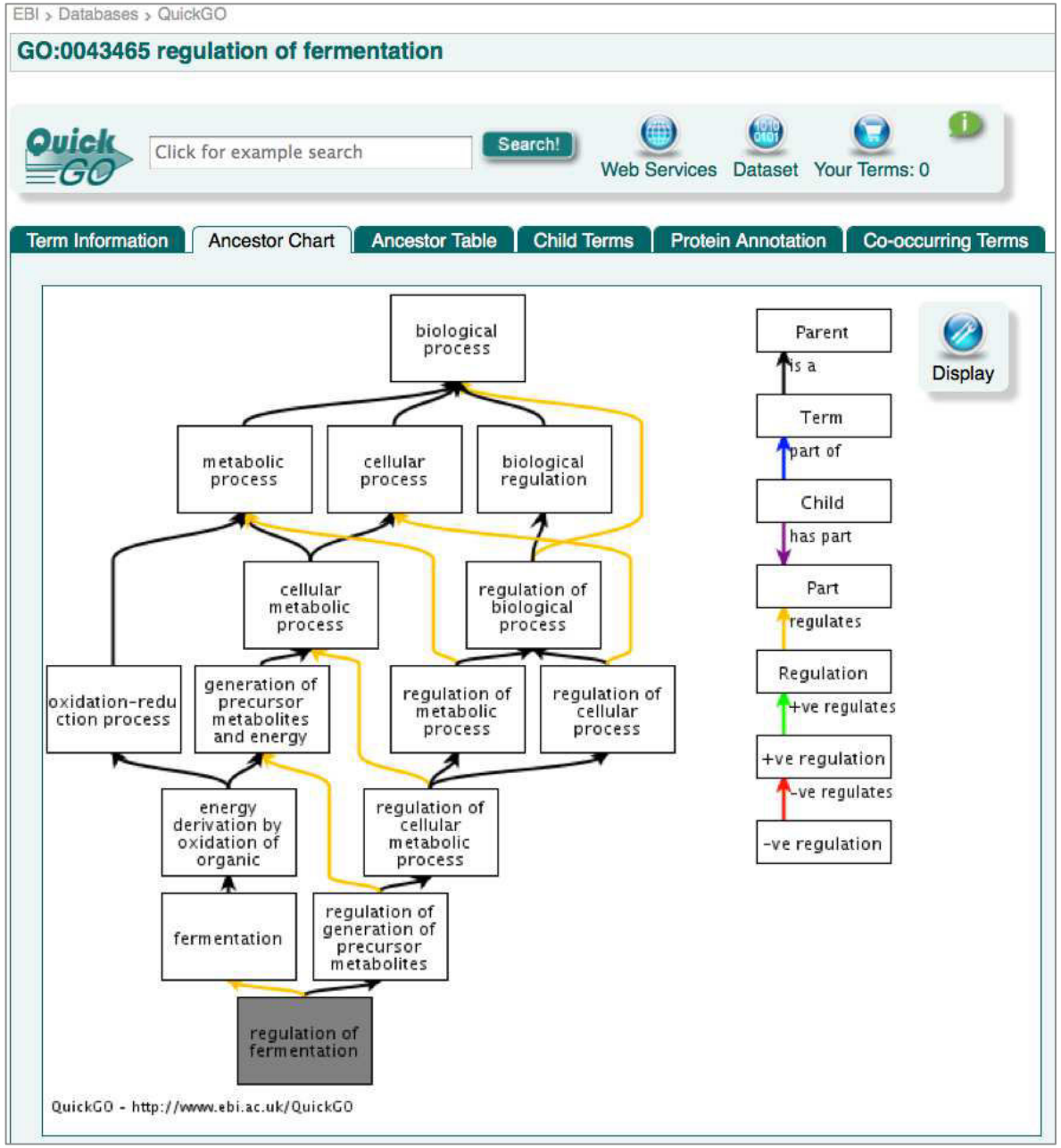
- Nombre de termes définis
  - Biological process 9,805
  - Molecular function 7,076
  - Cellular component 1,574
  - Sequence Ontology 963

- 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 convivale 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 ?

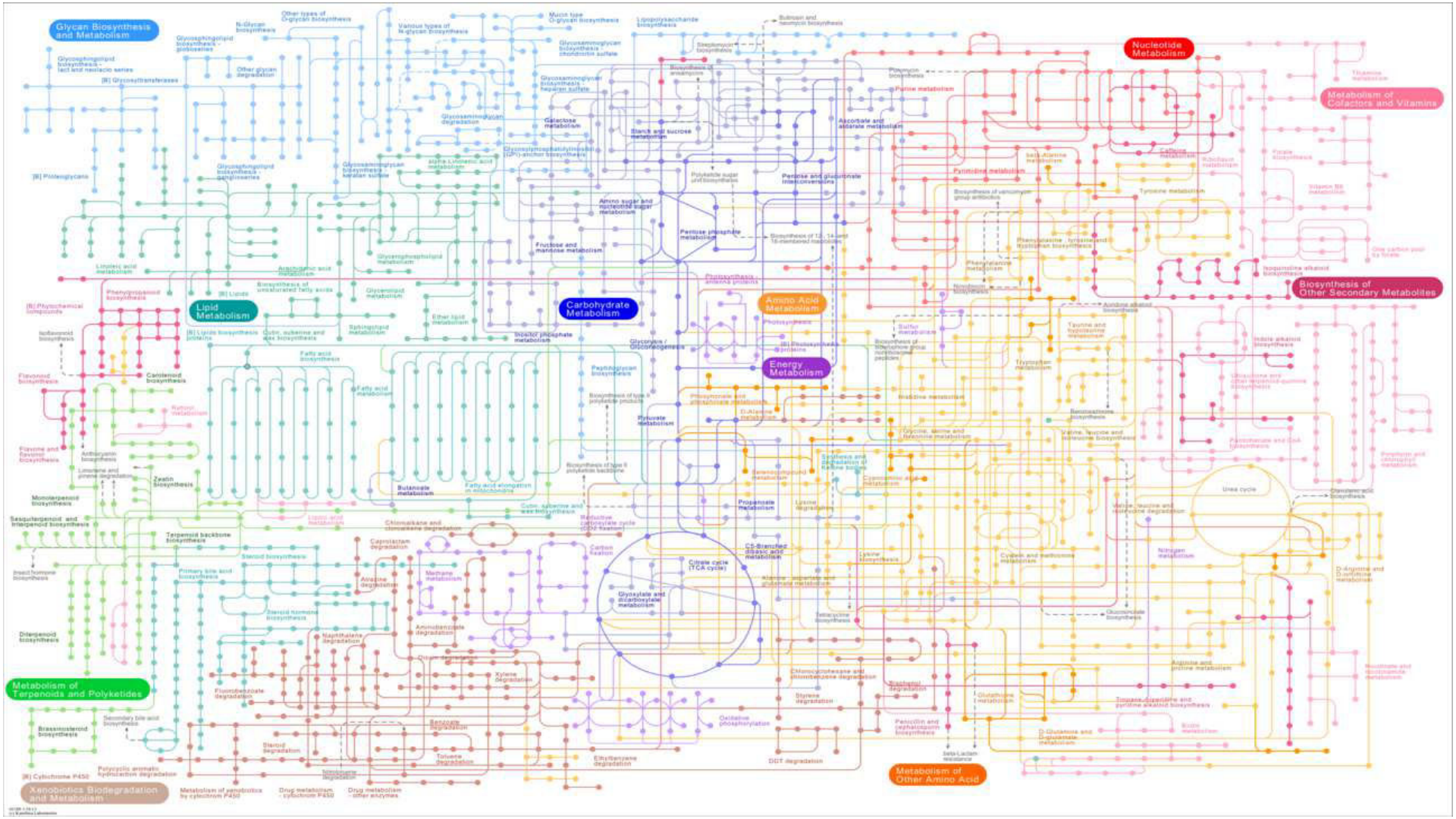
- A general definition
  - Fonction: action, rôle caractéristique d'un élément, d'un organe, dans un ensemble (souvent opposé à structure). Source: Le Petit Robert - dictionnaire alphabetique et analogique de la langue francaise. 1982.
- Représentation de la fonction dans la « gene ontology »
  - Dans 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: aspatokinase 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).

Petites molécules, réactions biochimiques et voies métaboliques

LIGAND – réactions métaboliques et petites molécules

KEGG - Kyoto Encycplopaedia of Genes and Genomes

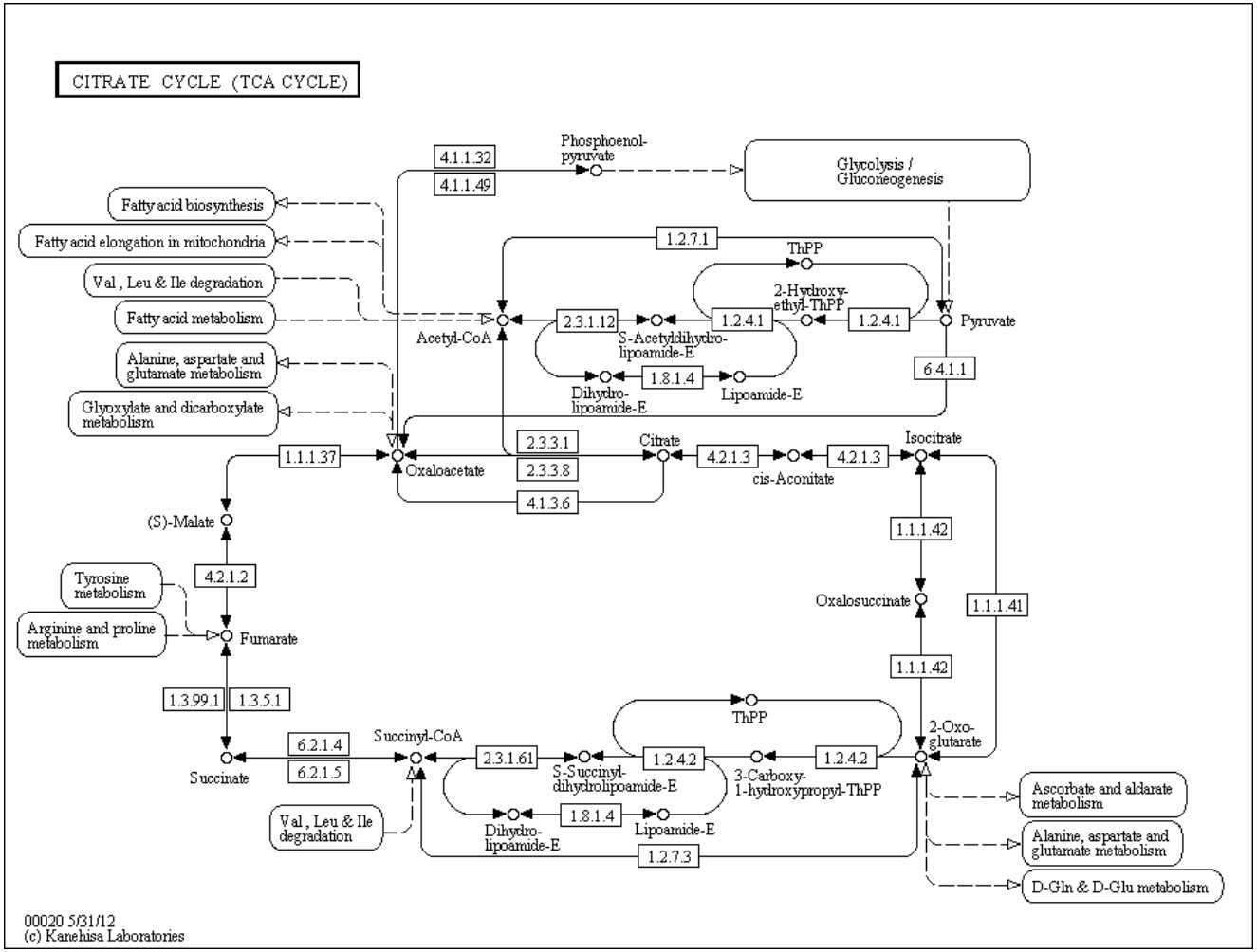
- La “carte globale” donne une vue d’ensemble de la complexité du métabolisme. Chaque point représente une molécule, chaque ligne une réaction métabolique.



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

KEGG - Kyoto Encycplopaedia 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 de biopuces

Ressources pour l'analyse du génome 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.