Cours en ligne

Paris 5

http://www.dsi.univ-paris5.fr/bio2/autof2/accueil2.htm

Pasteur

http://www.pasteur.fr/formation/infobio/infobio.html

ABI Marseille

http://www.up.univ-mrs.fr/wabim/biologie/tech.html

Gestion des informations génétiques

- Séquençage massif de chromosomes complets: information utilisable???
- Annotation des banques
- Structures en retard par rapport aux séquences
- Pas de banque totalisant toutes les informations sur une séquence

Banques de données

- Bases bibliographiques

- Bases de séquences
  - Généralistes
  - Spécialisées

- Bases de structures & domaines

- Banque = collecte des donnés exhaustive; hétérogènes
- Base = données homogènes (technique, espèce); valeur ajoutée

Banques de séquences

Acides nucléiques : Format commun depuis 1990

- **EMBL** (puis EBI) début 1980 : ADN
- **Genbank**, Los Alamos puis NCBI
- **DDBJ** (Japon)

Protéines :

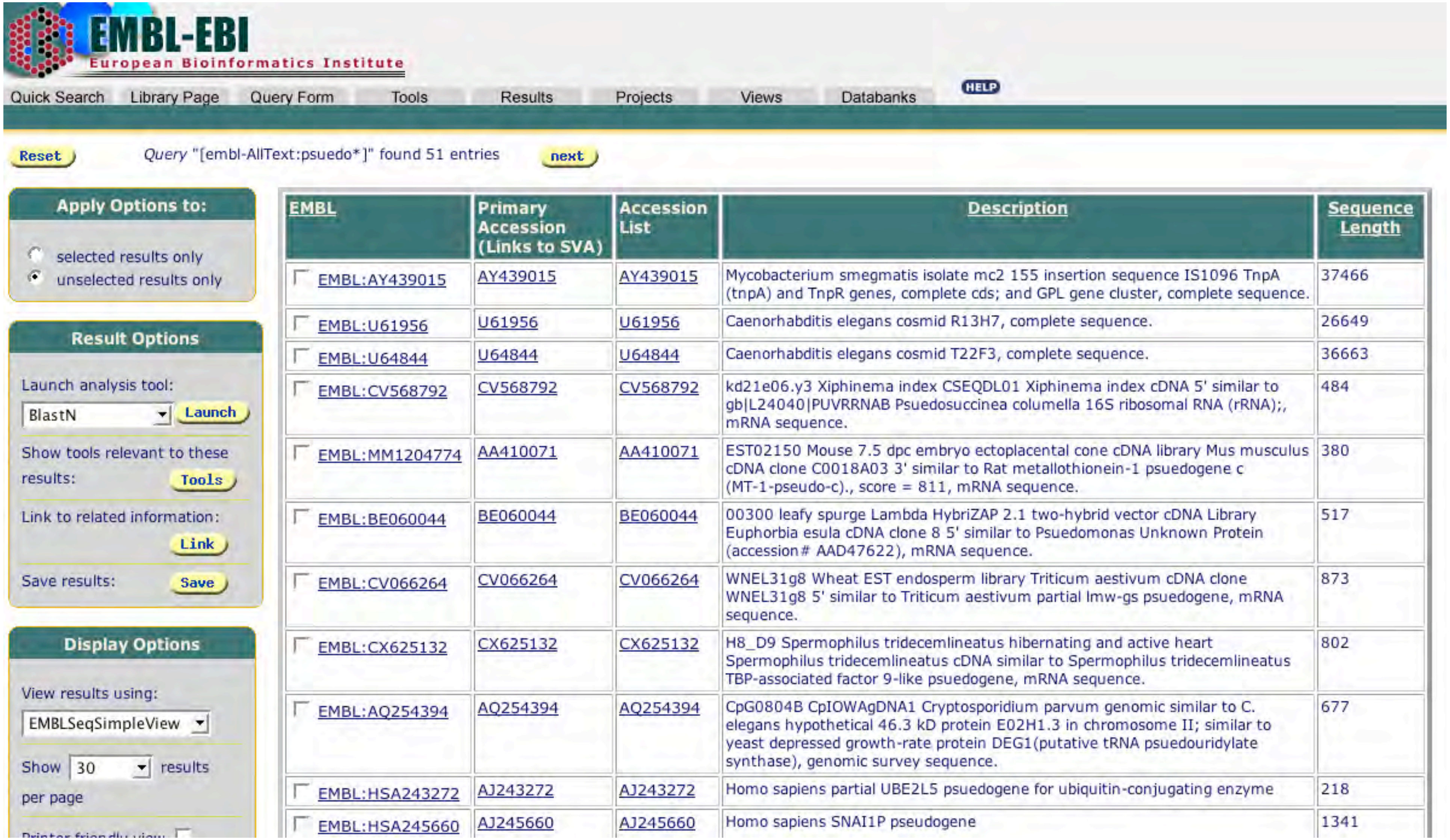
- **PIR** = NBRF (National Biomedical Research Foundation) + MIPS (Martinsried Institute for Protein Sequences) + JIPID (Japan International Protein Infromation Database)
- **SwissProt** = séquences annotées PIR-NBRF + séquences codantes traduites d'EMBL
- **Uniprot** = PIR+SP+TREMBL

Qualité données

- Information présente des lacunes
- Manque de vérifications systématiques des données soumises ou saisies, surtout pour les séquences anciennes
- Maintenant nombreuses vérifications systématiquement dès la soumission de la séquence
- On peut signaler aux responsables les erreurs dans les banques

Qualité données

Recherche de psuedo dans EMBL avec SRS



Query "[embl-AllText:psuedo*]" found 51 entries

○ selected results only
● unselected results only

Launch analysis tool: BlastN

Show tools relevant to these results:

Link to related information:

Save results:

View results using: EMBLSeqSimpleView

Show 30 results per page

EMBL	Primary Accession (Links to SVA)	Accession List	Description	Sequence Length
☐ EMBL:AY439015	AY439015	AY439015	Mycobacterium smegmatis isolate mc2 155 insertion sequence IS1096 TnpA (tnpA) and TnpR genes, complete cds; and GPL gene cluster, complete sequence.	37466
☐ EMBL:U61956	U61956	U61956	Caenorhabditis elegans cosmid R13H7, complete sequence.	26649
☐ EMBL:U64844	U64844	U64844	Caenorhabditis elegans cosmid T22F3, complete sequence.	36663
☐ EMBL:CV568792	CV568792	CV568792	kd21e06.y3 Xiphinema index CSEQDL01 Xiphinema index cDNA 5' similar to gb|L24040|PUVRRNAB Psuedosuccinea columella 16S ribosomal RNA (rRNA);, mRNA sequence.	484
☐ EMBL:MM1204774	AA410071	AA410071	EST02150 Mouse 7.5 dpc embryo ectoplacental cone cDNA library Mus musculus cDNA clone C0018A03 3' similar to Rat metallothionein-1 psuedogene c (MT-1-pseudo-c)., score = 811, mRNA sequence.	380
☐ EMBL:BE060044	BE060044	BE060044	00300 leafy spurge Lambda HybriZAP 2.1 two-hybrid vector cDNA Library Euphorbia esula cDNA clone 8 5' similar to Psuedomonas Unknown Protein (accession# AAD47622), mRNA sequence.	517
☐ EMBL:CV066264	CV066264	CV066264	WNEL31g8 Wheat EST endosperm library Triticum aestivum cDNA clone WNEL31g8 5' similar to Triticum aestivum partial lmw-gs psuedogene, mRNA sequence.	873
☐ EMBL:CX625132	CX625132	CX625132	H8_D9 Spermophilus tridecemlineatus hibernating and active heart Spermophilus tridecemlineatus cDNA similar to Spermophilus tridecemlineatus TBP-associated factor 9-like psuedogene, mRNA sequence.	802
☐ EMBL:AQ254394	AQ254394	AQ254394	CpG0804B CpIOWAgDNA1 Cryptosporidium parvum genomic similar to C. elegans hypothetical 46.3 kD protein E02H1.3 in chromosome II; similar to yeast depressed growth-rate protein DEG1(putative tRNA psuedouridylate synthase), genomic survey sequence.	677
☐ EMBL:HSA243272	AJ243272	AJ243272	Homo sapiens partial UBE2L5 psuedogene for ubiquitin-conjugating enzyme	218
☐ EMBL:HSA245660	AJ245660	AJ245660	Homo sapiens SNAI1P pseudogene	1341

Banques généralistes

Projets internationaux

- Indispensables à la communauté scientifique : données, résultats essentiels (certains ne sont plus dans la littérature, les noms changent)

- Principale mission = rendre publiques les séquences

- Bibliographie et expertise biologique liées aux séquences

- Organisation en rubriques ou champs; références croisées

- Pas de vérification pour les séquences anciennes

- Décalage entre dépôt et mise en ligne d'une séquence

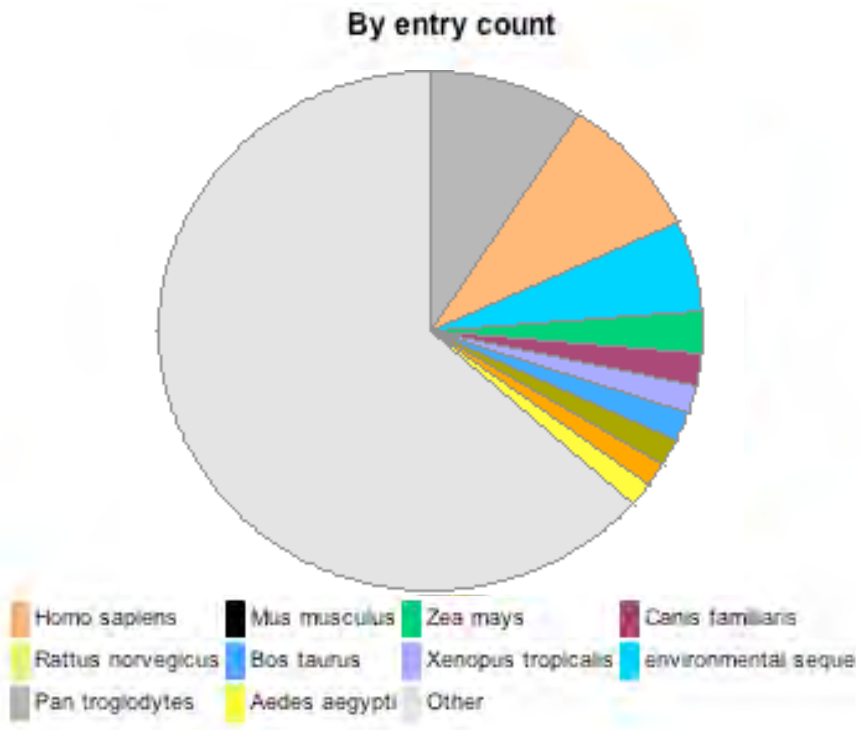
EMBL

Panda Group (Protein And Nucleotide Database) @ EBI

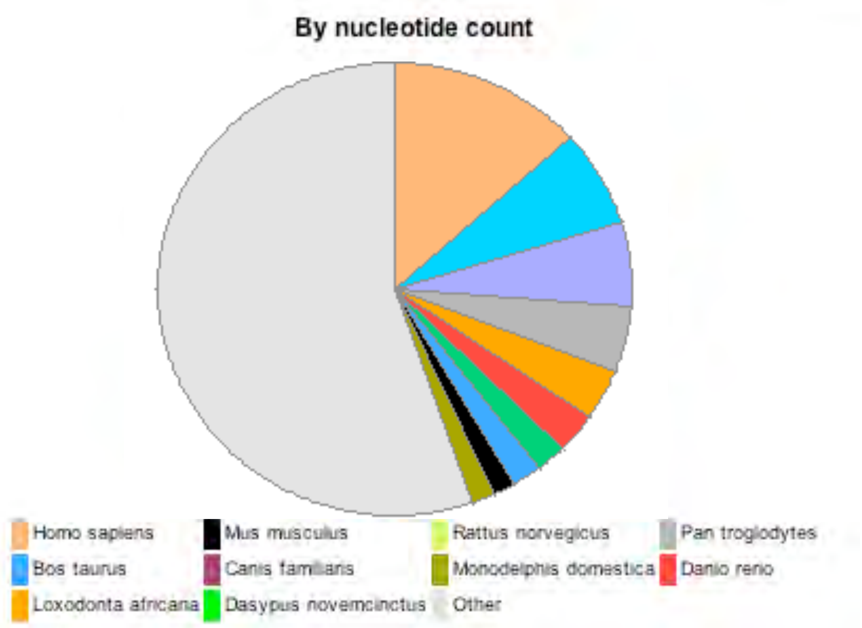
This morning the EMBL Database contained
236,378,259,855 nucleotides in **146,771,148** entries
Breakdown by entry type

Entry Type	**Entries**
Standard Constructed (CON)	2,595,559
Third Party Annotation (TPA)	6,008
Whole Genome Shotgun (WGS)	46,934,174

EMBL

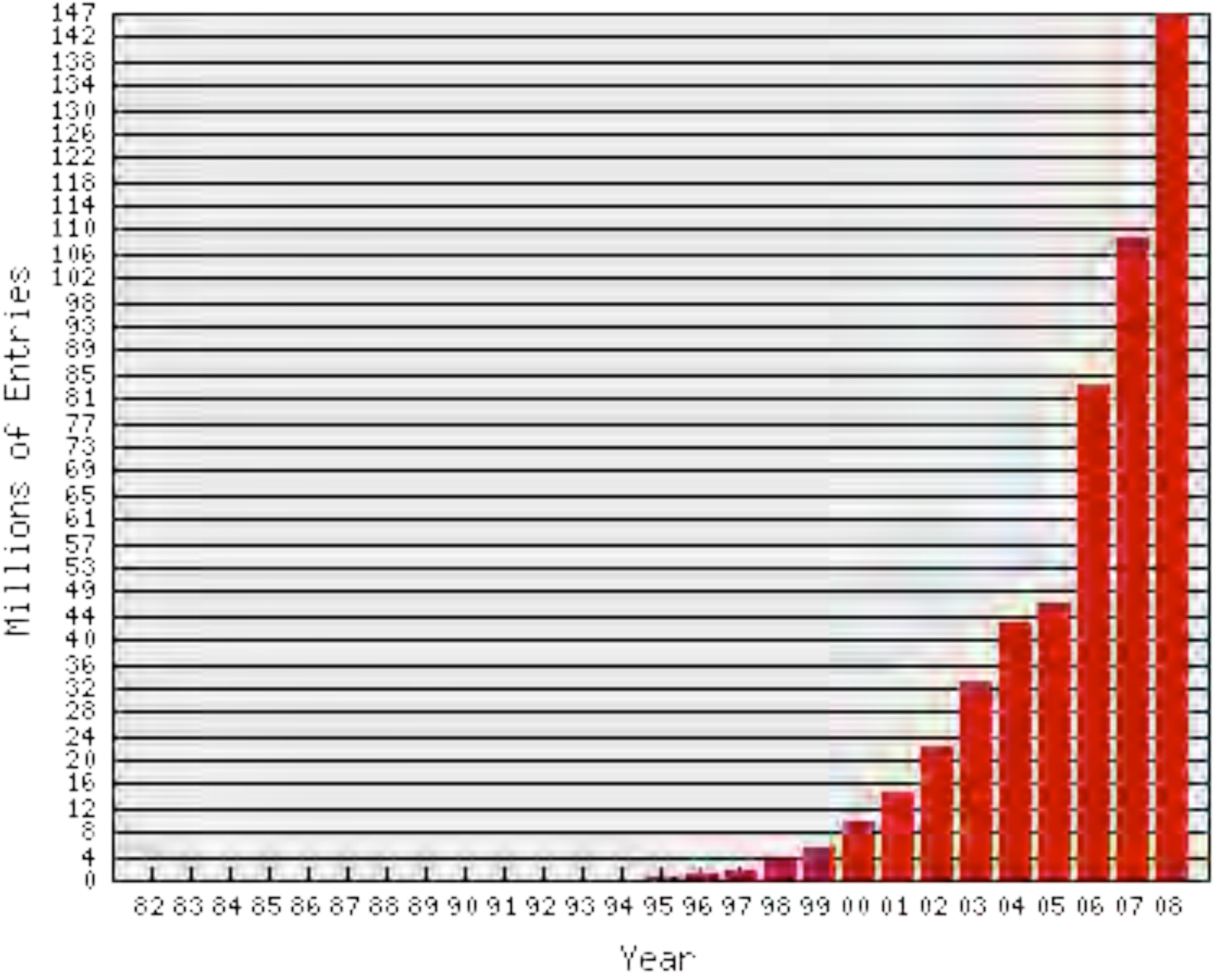


By entry count

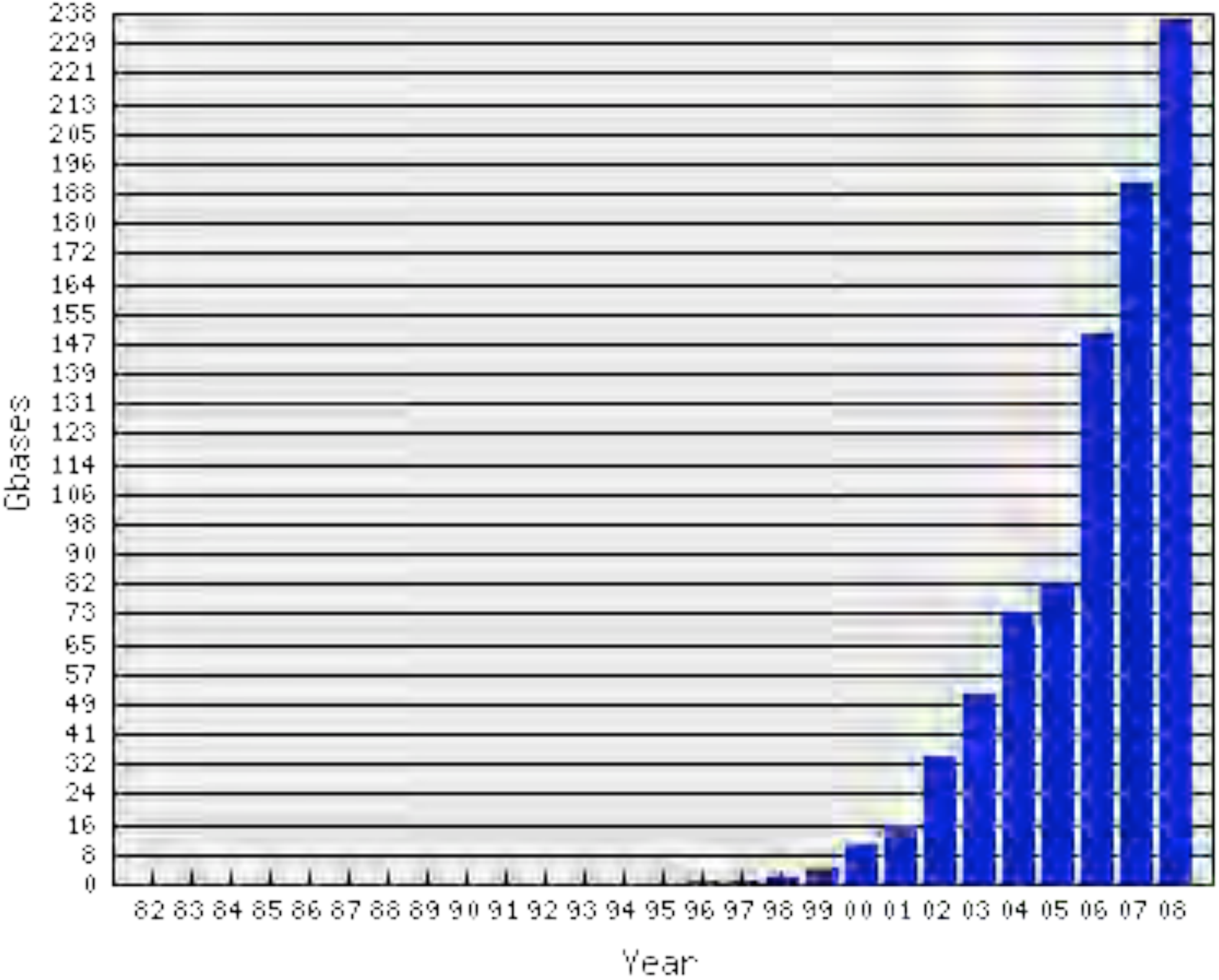


By nucleotide count

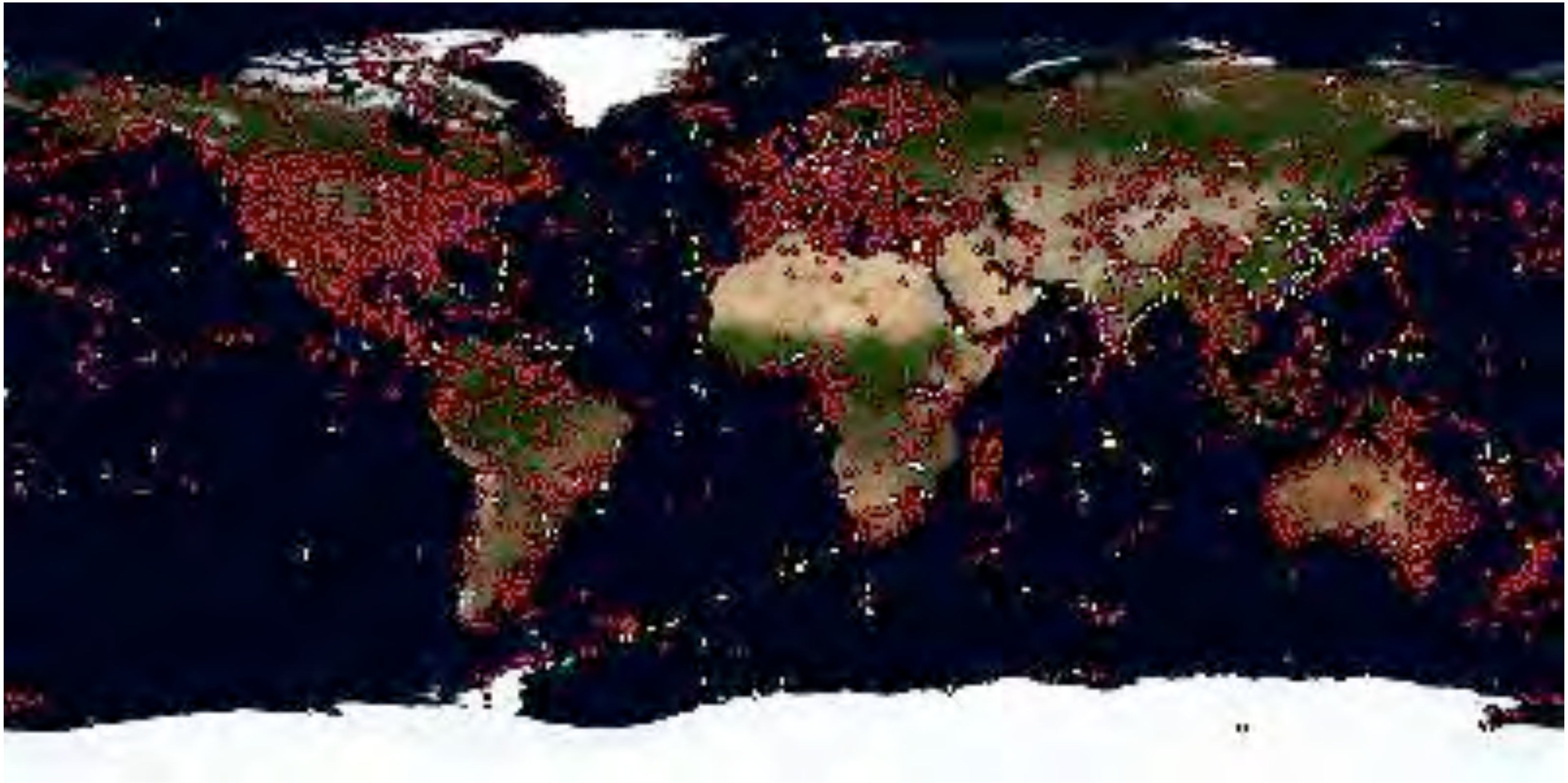
Croissance des entrées de la banque EMBL



EMBL: nombre de nucléotides

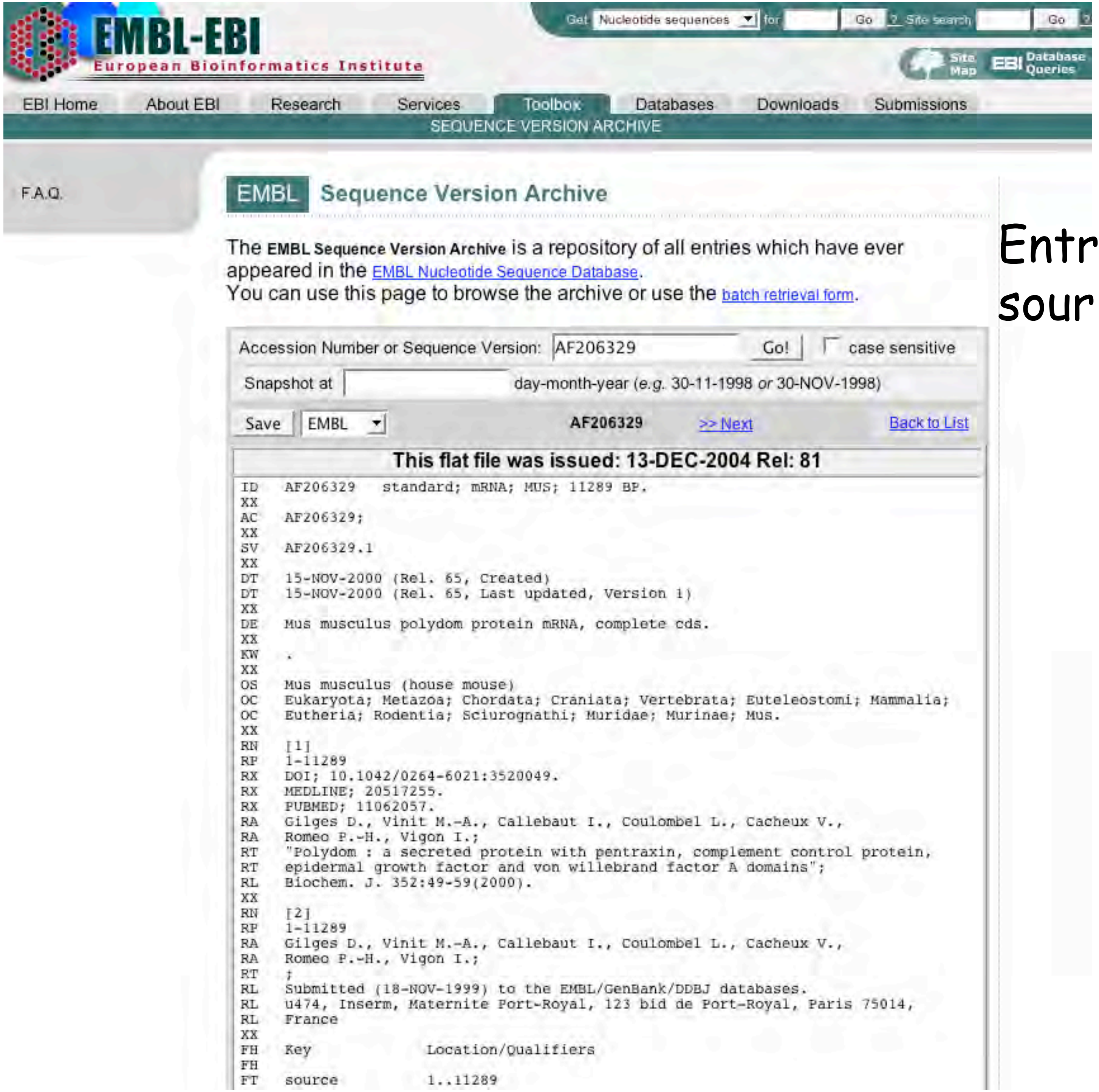


EMBL



EMBL

Entrée Polydom
souris



The **EMBL Sequence Version Archive** is a repository of all entries which have ever appeared in the EMBL Nucleotide Sequence Database.
You can use this page to browse the archive or use the batch retrieval form.

Accession Number or Sequence Version: AF206329 Go! ☐ case sensitive

Snapshot at ______ day-month-year (*e.g.* 30-11-1998 *or* 30-NOV-1998)

Save EMBL **AF206329** >> Next Back to List

This flat file was issued: 13-DEC-2004 Rel: 81

```
ID   AF206329   standard; mRNA; MUS; 11289 BP.
XX
AC   AF206329;
XX
SV   AF206329.1
XX
DT   15-NOV-2000 (Rel. 65, Created)
DT   15-NOV-2000 (Rel. 65, Last updated, Version 1)
XX
DE   Mus musculus polydom protein mRNA, complete cds.
XX
KW   .
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
XX
RN   [1]
RP   1-11289
RX   DOI; 10.1042/0264-6021:3520049.
RX   MEDLINE; 20517255.
RX   PUBMED; 11062057.
RA   Gilges D., Vinit M.-A., Callebaut I., Coulombel L., Cacheux V.,
RA   Romeo P.-H., Vigon I.;
RT   "Polydom : a secreted protein with pentraxin, complement control protein,
RT   epidermal growth factor and von willebrand factor A domains";
RL   Biochem. J. 352:49-59(2000).
XX
RN   [2]
RP   1-11289
RA   Gilges D., Vinit M.-A., Callebaut I., Coulombel L., Cacheux V.,
RA   Romeo P.-H., Vigon I.;
RT   ;
RL   Submitted (18-NOV-1999) to the EMBL/GenBank/DDBJ databases.
RL   u474, Inserm, Maternite Port-Royal, 123 bid de Port-Royal, Paris 75014,
RL   France
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..11289
```

EMBL

```
KW   .
XX
OS   Mus musculus (house mouse)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
XX
RN   [1]
RP   1-11289
RX   DOI; 10.1042/0264-6021:3520049.
RX   MEDLINE; 20517255.
RX   PUBMED; 11062057.
RA   Gilges D., Vinit M.-A., Callebaut I., Coulombel L., Cacheux V.,
RA   Romeo P.-H., Vigon I.;
RT   "Polydom : a secreted protein with pentraxin, complement control protein,
RT   epidermal growth factor and von willebrand factor A domains";
RL   Biochem. J. 352:49-59(2000).
XX
RN   [2]
RP   1-11289
RA   Gilges D., Vinit M.-A., Callebaut I., Coulombel L., Cacheux V.,
RA   Romeo P.-H., Vigon I.;
RT   ;
RL   Submitted (18-NOV-1999) to the EMBL/GenBank/DDBJ databases.
RL   u474, Inserm, Maternite Port-Royal, 123 bld de Port-Royal, Paris 75014,
RL   France
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..11289
FT                   /chromosome="4"
FT                   /db_xref="taxon:10090"
FT                   /mol_type="mRNA"
FT                   /organism="Mus musculus"
FT                   /strain="C3H/HeNSIc"
FT                   /map="4b-4c2"
FT                   /cell_line="MS-5 stromal cells"
FT   sig_peptide     221..271
FT   CDS             221..10924
FT                   /codon_start=1
FT                   /db_xref="GOA:Q9ES77"
FT                   /db_xref="HSSP:Q9UCV4"
FT                   /db_xref="MGI:1928849"
FT                   /db_xref="UniProt/TrEMBL:Q9ES77"
FT                   /product="polydom protein"
FT                   /protein_id="AAG32160.1"
FT                   /translation="MWSRLAFCCWALALVSGWTNFQPVAPSLNFSFRLFPEASPGALGR
FT                   LAVPPASSEEEAAGSKVERLGRAFRSRVRRLRELSGSLELVFLVDESSSVGQTNFLNEL
FT                   KFVRKLLSDFPVVSTATRVAIVTFSSKNNVVARVDYISTSRAHQHKCALLSREIPAITY
FT                   RGGGTYTKGAFQQAAQILRHSRENSTKVIFLITDGYSNGGDPRPIAASLRDFGVEIFTF
FT                   GIWQGNIRELNDMASTPKEEHCYLLHSFEEFEALARRALHEDLPSGSFIQEDMARCSYL
FT                   CEAGKDCCDRMASCKCGTHTGQFECICEKGYYGKGLQHECTACPSGTYKPEASPGGIST
FT                   CIPCPDVSHTSPPGSTSPEDCVCREGYQRSGQTCEVVHCPALKPPENGFFIQNTCKNHF
FT                   NAACGVRCRPGFDLVGSSIHLCQPNGLWSGTESFCRVRTCPHLRQPKHGHISCSTAEMS
```

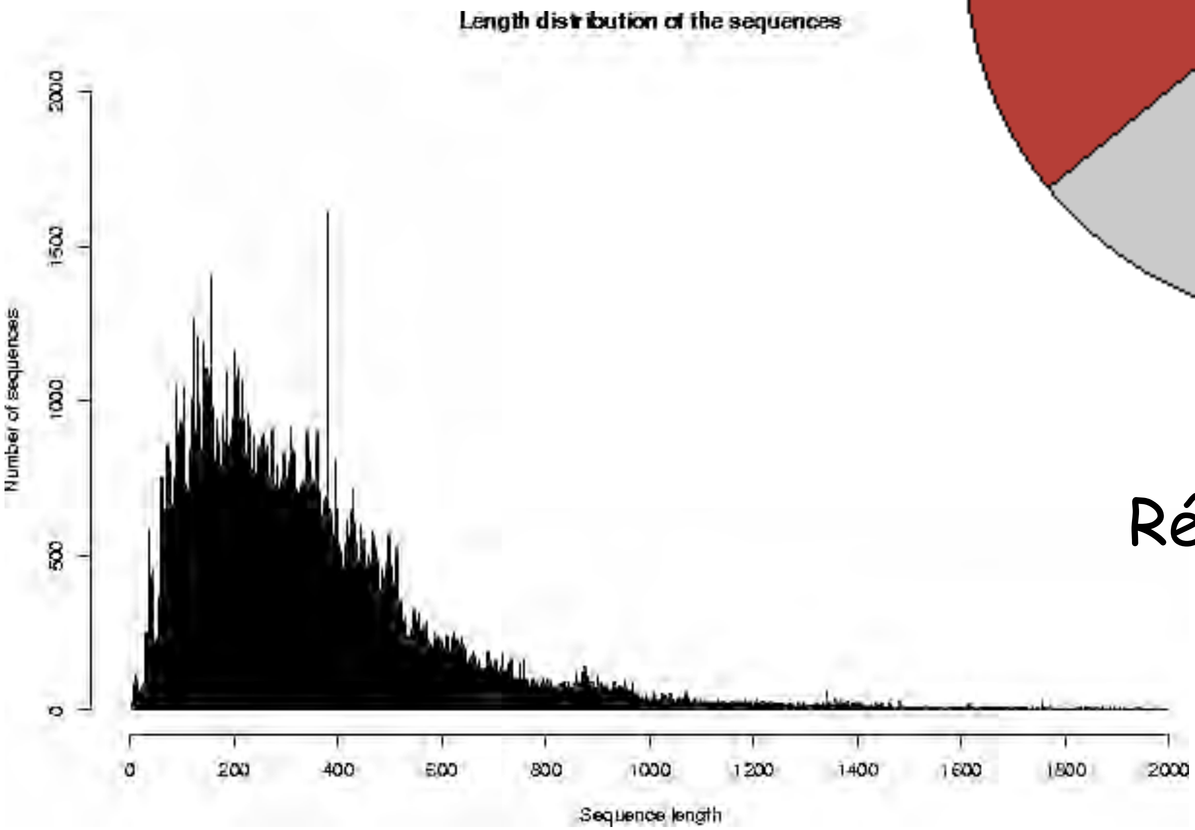
SWISS-PROT

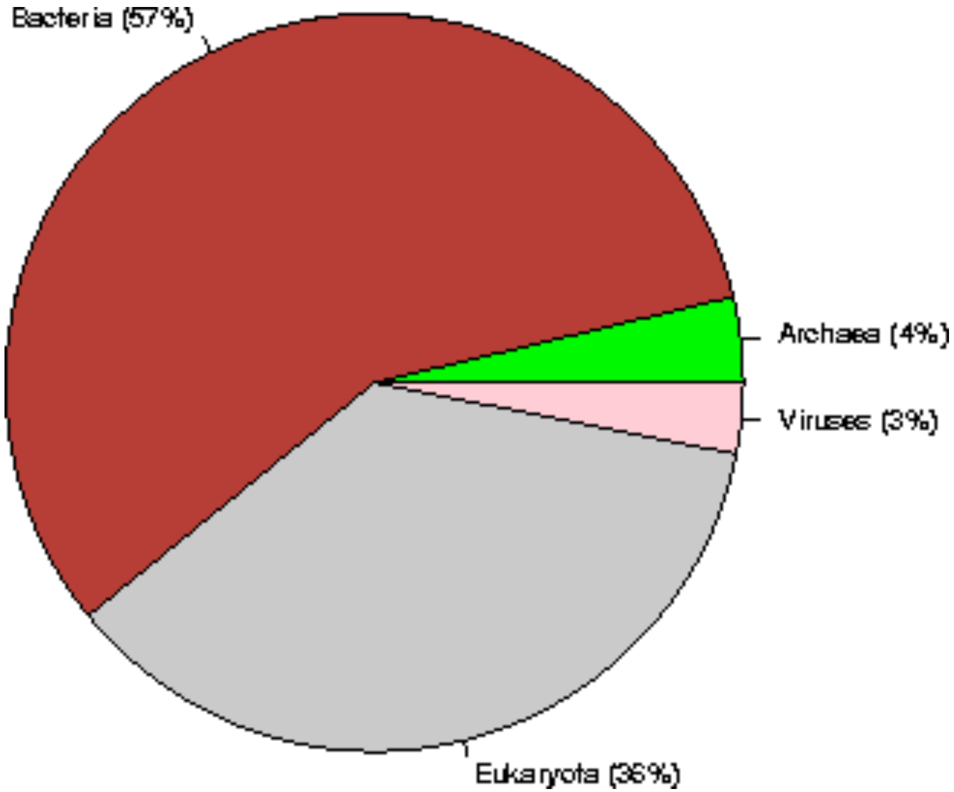
- 1986. Amos Bairoch. Department of Medical Biochemistry of the University of Geneva & EMBL

- Release 56.3 of 14-Oct-08 of UniProtKB/Swiss-Prot contains 399 749 sequence entries, comprising 144 041 553 amino acids abstracted from 173629 references

SWISS-PROT

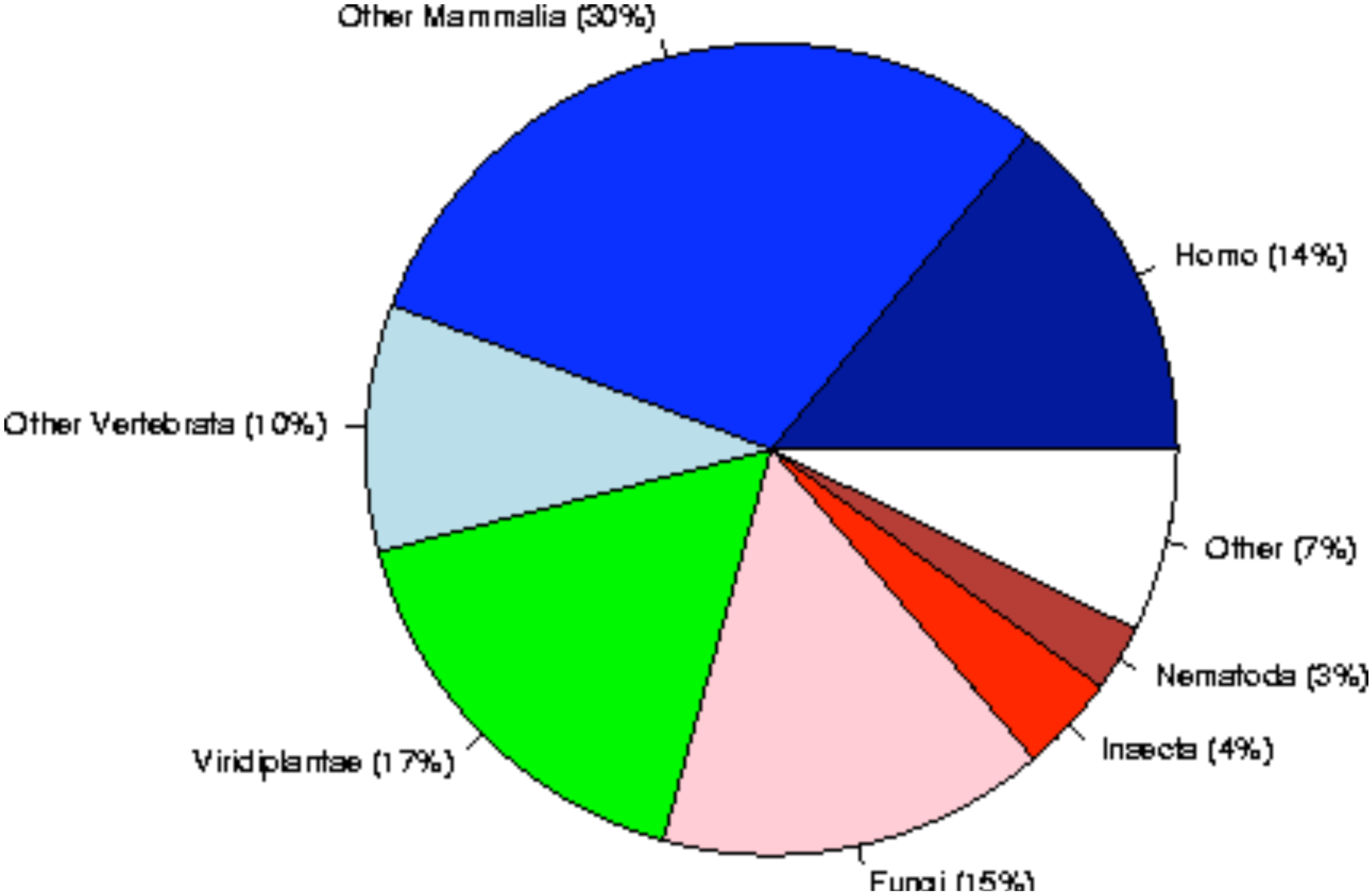
Distribution des longueurs



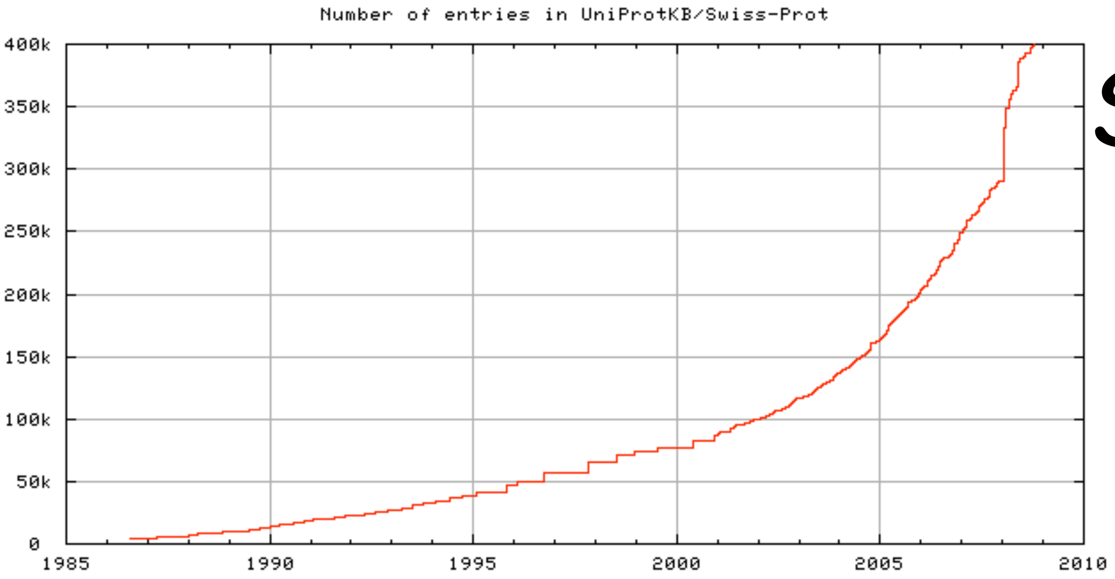


Répartition des entrées

SWISS-PROT

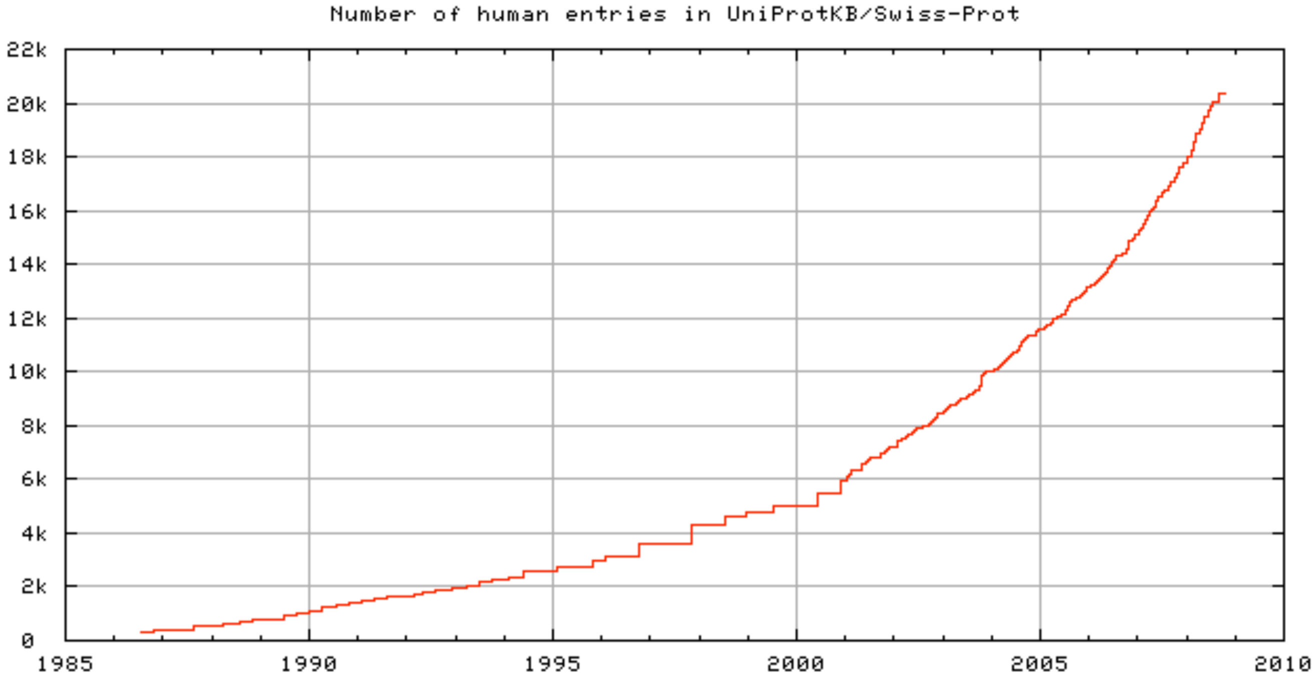

Other Mammalia (30%)
Homo (14%)
Other Vertebrata (10%)
Other (7%)
Nematoda (3%)
Insecta (4%)
Viridiplantae (17%)
Fungi (15%)

SWISS-PROT



Nombre d'entrées dans SP

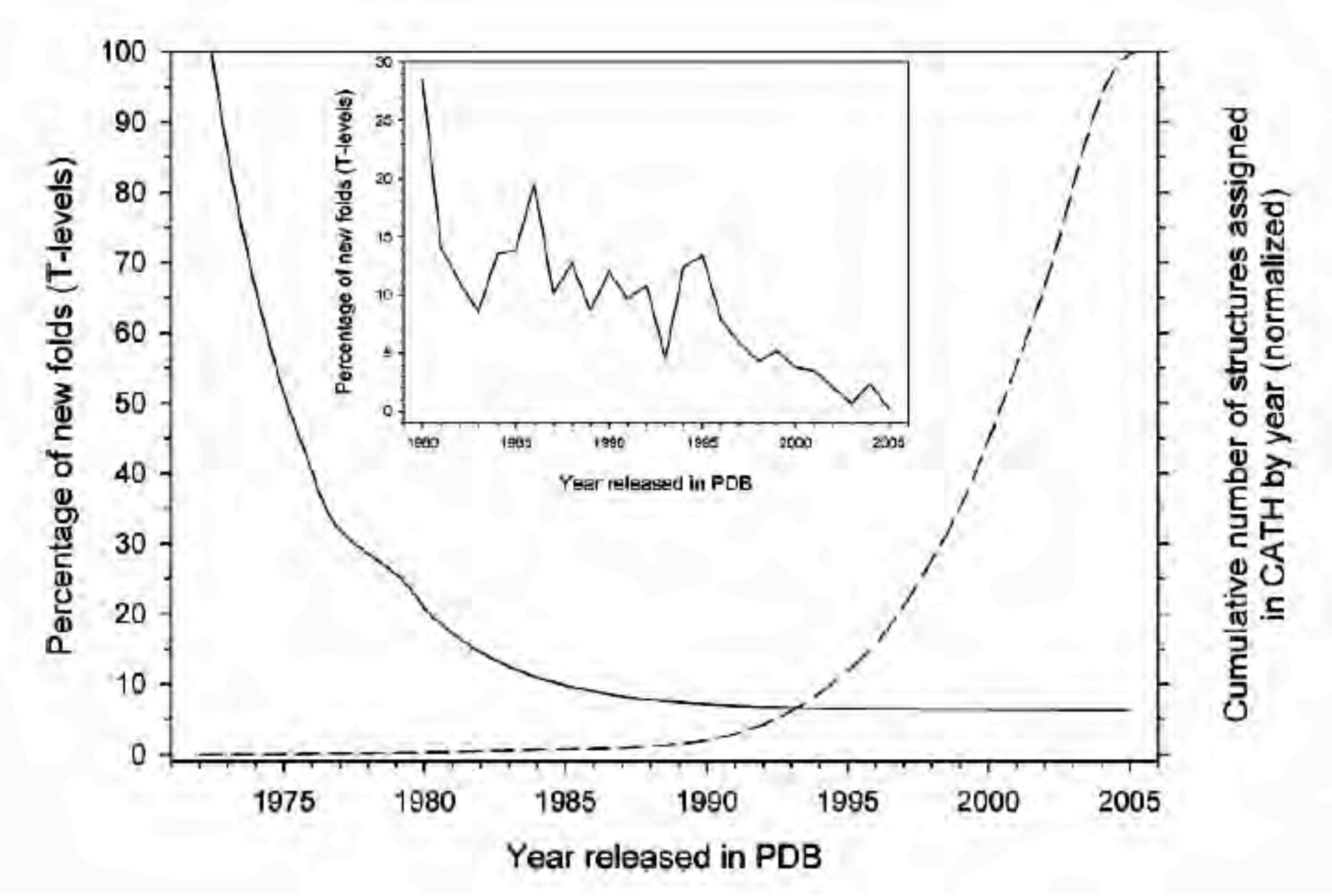
Nombre d'entrées humaines dans SP



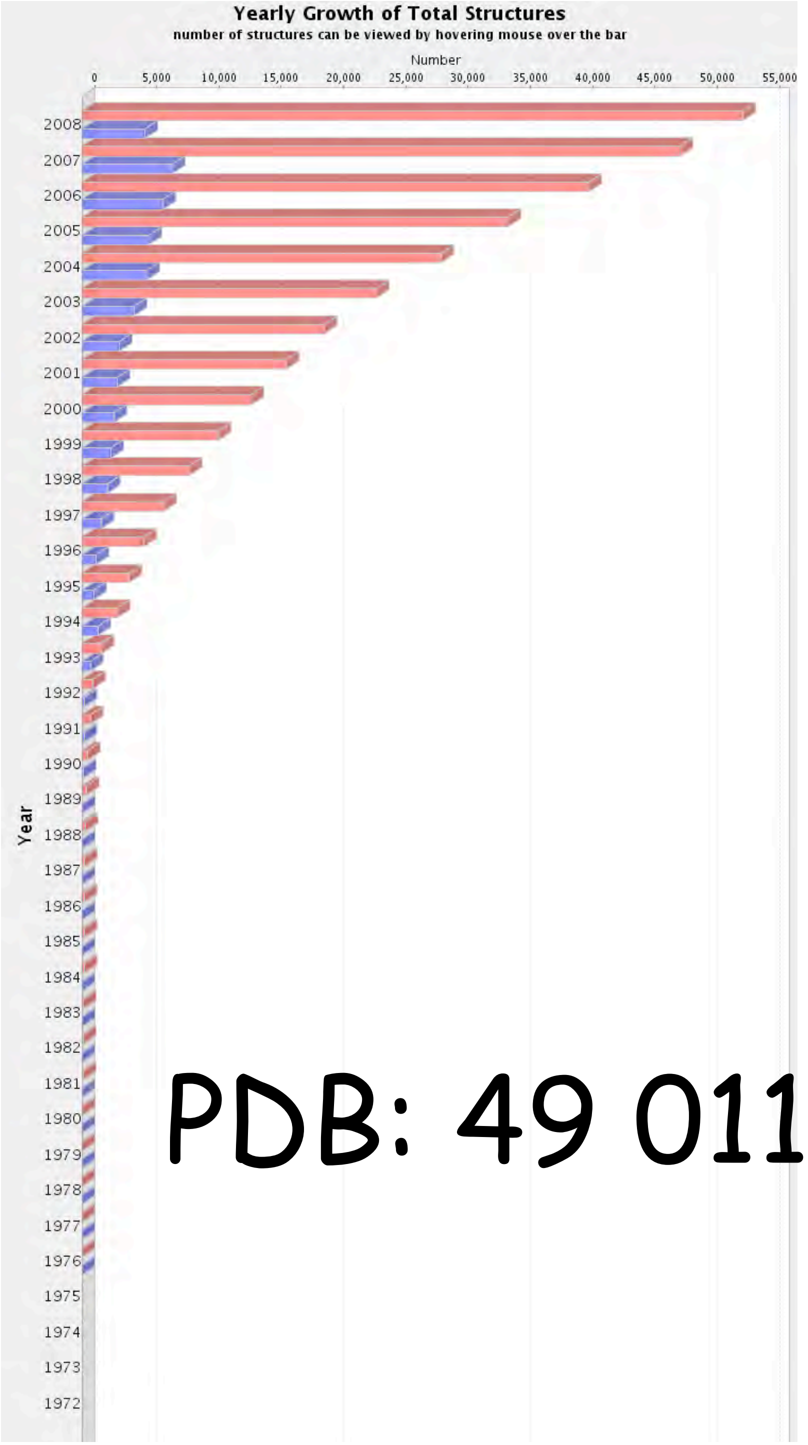
SWISS-PROT

Croissance des banques de séquence beaucoup plus rapide que celle des banques de structures

Taux de nouveaux repliements



PDB


Yearly Growth of Total Structures

number of structures can be viewed by hovering mouse over the bar


Growth Of Unique Folds Per Year As Defined By SCOP

number of folds can be viewed by hovering mouse over the bar

PDB: 49 011 protéines en 2008

SWISS -PROT

Swiss-Prot
Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:

- **Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].
- **TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Release 4.2 consists of:
Swiss-Prot Release 46.2 of 01-Mar-2005: 172233 entries
(More statistics)
TrEMBL Release 29.2 of 01-Mar-2005: 1631173 entries
(More statistics)

> Swiss-Prot headlines
More than 10'000 additional sequences encoded on splice variants in Swiss-Prot (Read more...)

Access to Swiss-Prot and TrEMBL

- SRS - Access to Swiss-Prot, TrEMBL and other databases using the Sequence Retrieval System
- Full text search in Swiss-Prot and TrEMBL
- Advanced search in Swiss-Prot and TrEMBL by description, gene name and organism (can be used to create html links to Swiss-Prot/TrEMBL queries)
- Taxonomy browser (NEWT)

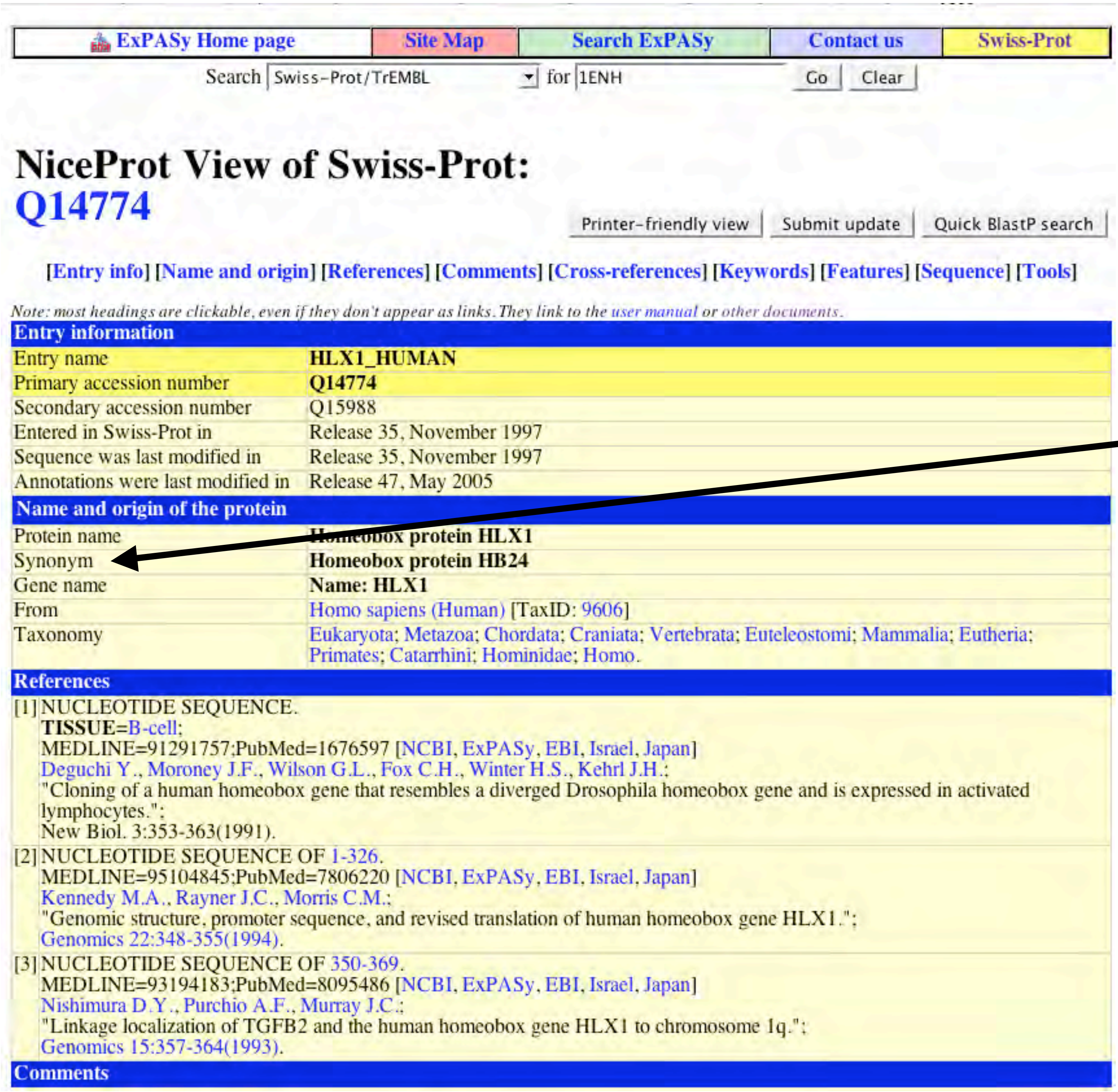
- by description or identification (any word in the DE, OS, OG, GN and ID lines; Swiss-Prot and TrEMBL)
- by citation (RL line; Swiss-Prot only)

- Retrieve a list of Swiss-Prot/TrEMBL entries
- Randomly retrieve a Swiss-Prot/TrEMBL entry
- Swiss-Prot ID tracker

Documents and services

- Swiss-Prot documents - user manual, release notes, indices and lots of other **important** documents and lists
- Swiss-Shop - a service that allows you to automatically obtain (by email) new sequence entries relevant to your field(s) of interest

SWISS -PROT



NiceProt View of Swiss-Prot: Q14774

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**HLX1_HUMAN**
Primary accession number	**Q14774**
Secondary accession number	Q15988
Entered in Swiss-Prot in	Release 35, November 1997
Sequence was last modified in	Release 35, November 1997
Annotations were last modified in	Release 47, May 2005
Name and origin of the protein	
Protein name	**Homeobox protein HLX1**
Synonym	**Homeobox protein HB24**
Gene name	**Name: HLX1**
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

References

[1] NUCLEOTIDE SEQUENCE.
TISSUE=B-cell;
MEDLINE=91291757;PubMed=1676597 [NCBI, ExPASy, EBI, Israel, Japan]
Deguchi Y., Moroney J.F., Wilson G.L., Fox C.H., Winter H.S., Kehrl J.H.;
"Cloning of a human homeobox gene that resembles a diverged Drosophila homeobox gene and is expressed in activated lymphocytes.";
New Biol. 3:353-363(1991).

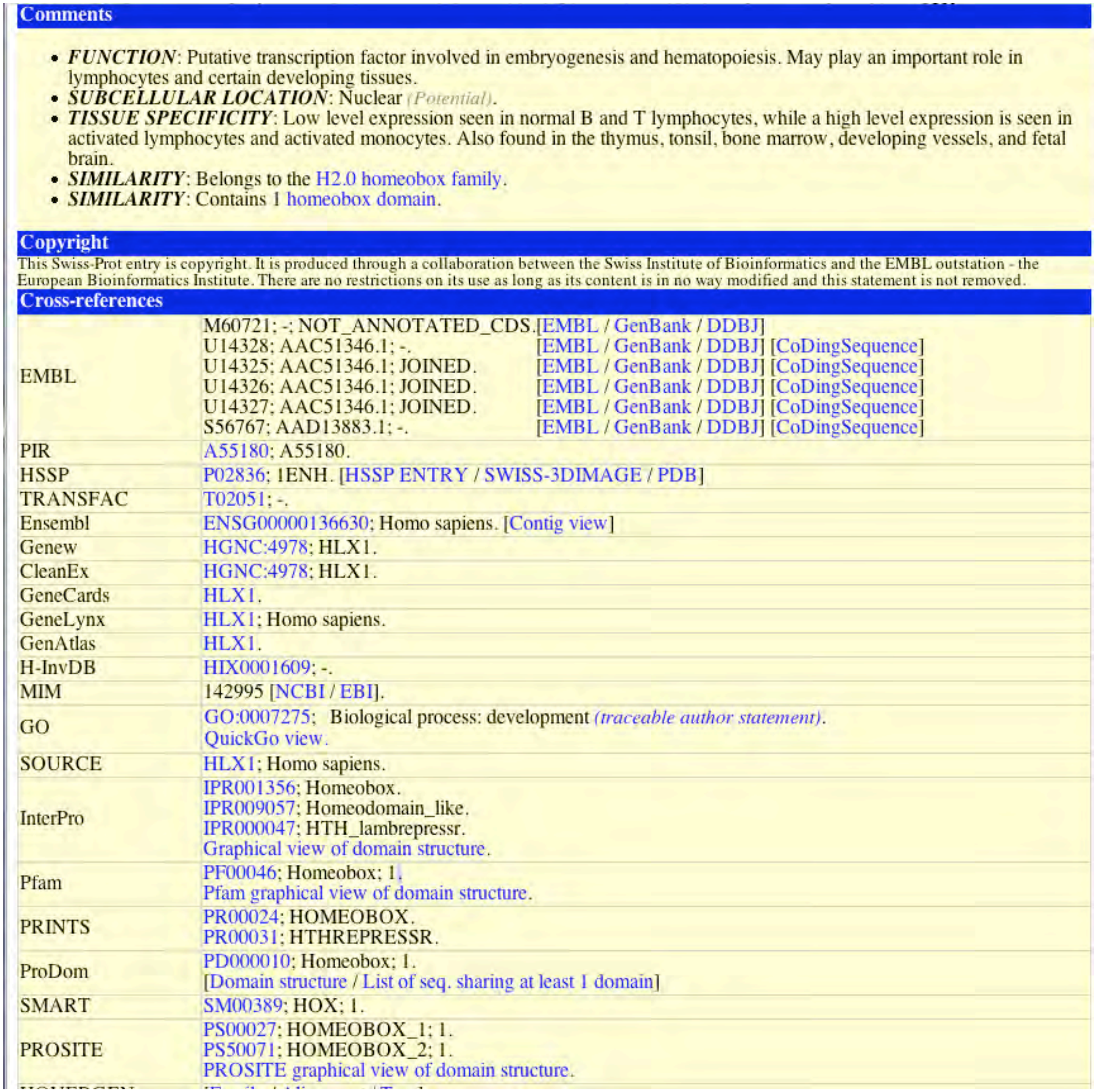
[2] NUCLEOTIDE SEQUENCE OF 1-326.
MEDLINE=95104845;PubMed=7806220 [NCBI, ExPASy, EBI, Israel, Japan]
Kennedy M.A., Rayner J.C., Morris C.M.;
"Genomic structure, promoter sequence, and revised translation of human homeobox gene HLX1.";
Genomics 22:348-355(1994).

[3] NUCLEOTIDE SEQUENCE OF 350-369.
MEDLINE=93194183;PubMed=8095486 [NCBI, ExPASy, EBI, Israel, Japan]
Nishimura D.Y., Purchio A.F., Murray J.C.;
"Linkage localization of TGFB2 and the human homeobox gene HLX1 to chromosome 1q.";
Genomics 15:357-364(1993).

Comments

Plusieurs noms possibles

SWISS -PROT



Comments

- ***FUNCTION***: Putative transcription factor involved in embryogenesis and hematopoiesis. May play an important role in lymphocytes and certain developing tissues.
- ***SUBCELLULAR LOCATION***: Nuclear *(Potential)*.
- ***TISSUE SPECIFICITY***: Low level expression seen in normal B and T lymphocytes, while a high level expression is seen in activated lymphocytes and activated monocytes. Also found in the thymus, tonsil, bone marrow, developing vessels, and fetal brain.
- ***SIMILARITY***: Belongs to the H2.0 homeobox family.
- ***SIMILARITY***: Contains 1 homeobox domain.

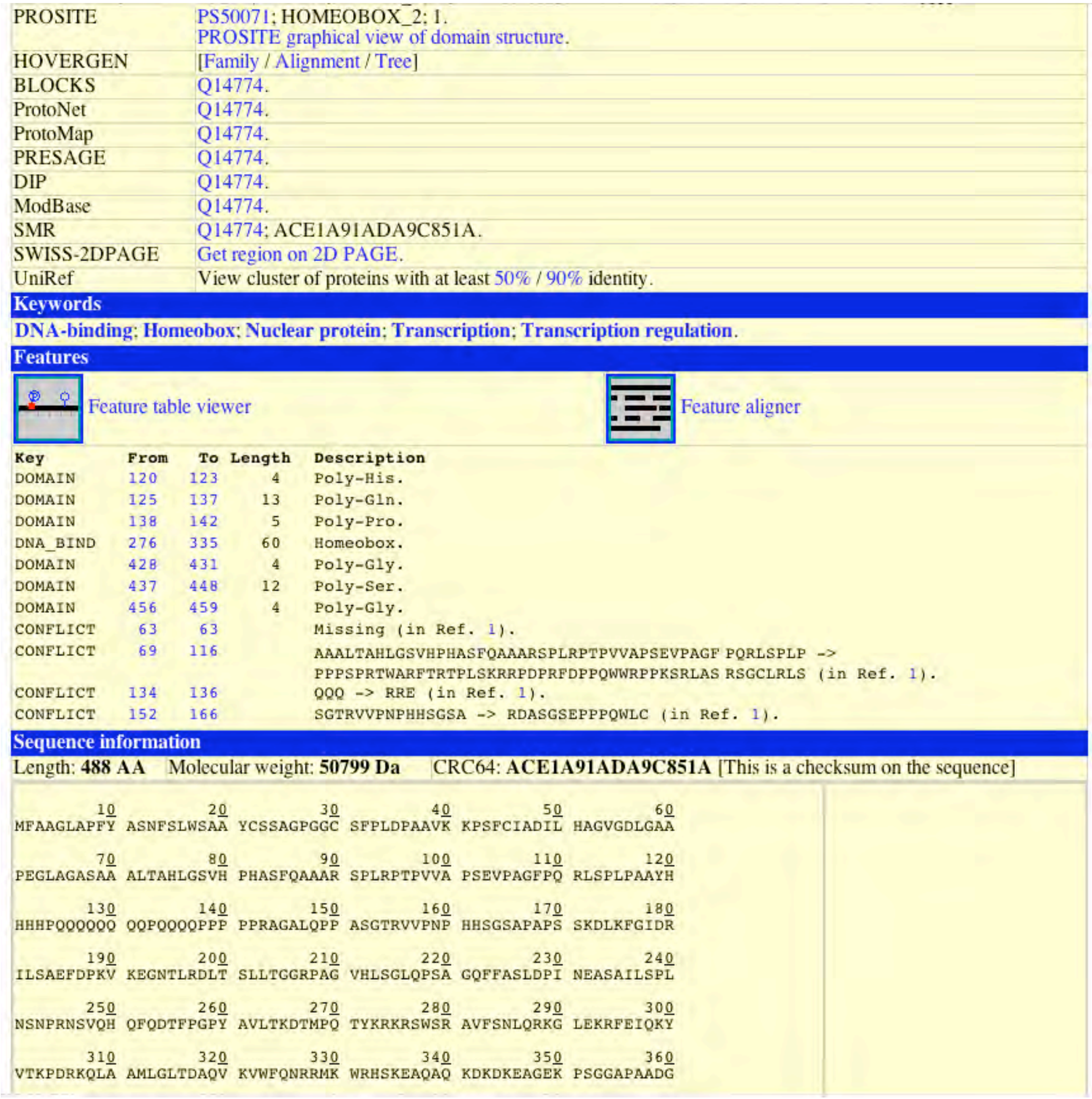
Copyright

This Swiss-Prot entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use as long as its content is in no way modified and this statement is not removed.

Cross-references

EMBL	M60721; -; NOT_ANNOTATED_CDS. [EMBL / GenBank / DDBJ] U14328; AAC51346.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] U14325; AAC51346.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] U14326; AAC51346.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] U14327; AAC51346.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] S56767; AAD13883.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A55180; A55180.
HSSP	P02836; 1ENH. [HSSP ENTRY / SWISS-3DIMAGE / PDB]
TRANSFAC	T02051; -.
Ensembl	ENSG00000136630; Homo sapiens. [Contig view]
Genew	HGNC:4978; HLX1.
CleanEx	HGNC:4978; HLX1.
GeneCards	HLX1.
GeneLynx	HLX1; Homo sapiens.
GenAtlas	HLX1.
H-InvDB	HIX0001609; -.
MIM	142995 [NCBI / EBI].
GO	GO:0007275; Biological process: development *(traceable author statement)*. QuickGo view.
SOURCE	HLX1; Homo sapiens.
InterPro	IPR001356; Homeobox. IPR009057; Homeodomain_like. IPR000047; HTH_lambrepressr. Graphical view of domain structure.
Pfam	PF00046; Homeobox; 1. Pfam graphical view of domain structure.
PRINTS	PR00024; HOMEOBOX. PR00031; HTHREPRESSR.
ProDom	PD000010; Homeobox; 1. [Domain structure / List of seq. sharing at least 1 domain]
SMART	SM00389; HOX; 1.
PROSITE	PS00027; HOMEOBOX_1; 1. PS50071; HOMEOBOX_2; 1. PROSITE graphical view of domain structure.

SWISS -PROT



PROSITE	PS50071; HOMEOBOX_2; 1. PROSITE graphical view of domain structure.
HOVERGEN	[Family / Alignment / Tree]
BLOCKS	Q14774.
ProtoNet	Q14774.
ProtoMap	Q14774.
PRESAGE	Q14774.
DIP	Q14774.
ModBase	Q14774.
SMR	Q14774; ACE1A91ADA9C851A.
SWISS-2DPAGE	Get region on 2D PAGE.
UniRef	View cluster of proteins with at least 50% / 90% identity.

Keywords

DNA-binding; **Homeobox**; **Nuclear protein**; **Transcription**; **Transcription regulation**.

Features

```
Key           From   To Length   Description
DOMAIN         120  123      4   Poly-His.
DOMAIN         125  137     13   Poly-Gln.
DOMAIN         138  142      5   Poly-Pro.
DNA_BIND       276  335     60   Homeobox.
DOMAIN         428  431      4   Poly-Gly.
DOMAIN         437  448     12   Poly-Ser.
DOMAIN         456  459      4   Poly-Gly.
CONFLICT        63   63          Missing (in Ref. 1).
CONFLICT        69  116          AAALTAHLGSVHPHASFQAAARSPLRPTPVVAPSEVPAGF PQRLSPLP ->
                                 PPPSPRTWARFTRTPLSKRRPDPRFDPPQWWRPPKSRLAS RSGCLRLS (in Ref. 1).
CONFLICT       134  136          QQQ -> RRE (in Ref. 1).
CONFLICT       152  166          SGTRVVPNPHHSGSA -> RDASGSEPPPQWLC (in Ref. 1).
```

Sequence information

Length: **488 AA** Molecular weight: **50799 Da** CRC64: **ACE1A91ADA9C851A** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MFAAGLAPFY ASNFSLWSAA YCSSAGPGGC SFPLDPAAVK KPSFCIADIL HAGVGDLGAA

        70         80         90        100        110        120
PEGLAGASAA ALTAHLGSVH PHASFQAAAR SPLRPTPVVA PSEVPAGFPQ RLSPLPAAYH

       130        140        150        160        170        180
HHHPQQQQQQ QQPQQQQPPP PPRAGALQPP ASGTRVVPNP HHSGSAPAPS SKDLKFGIDR

       190        200        210        220        230        240
ILSAEFDPKV KEGNTLRDLT SLLTGGRPAG VHLSGLQPSA GQFFASLDPI NEASAILSPL

       250        260        270        280        290        300
NSNPRNSVQH QFQDTFPGPY AVLTKDTMPQ TYKRKRSWSR AVFSNLQRKG LEKRFEIQKY

       310        320        330        340        350        360
VTKPDRKQLA AMLGLTDAQV KVWFQNRRMK WRHSKEAQAQ KDKDKEAGEK PSGGAPAADG
```

Noms

- En biologie, les noms des protéines ne sont pas stabilisés: évoluent avec le temps

- Une protéine peut avoir plusieurs fonctions

Synonymes Interleukine 6

- D'après Robert Nitsch
- Interferon β2
- B cell differentiation factor
- B cell stimulatory factor 2
- Hepatocyte stimulatory factor
- Hibridoma growth factor
- Plasmocytoma growth factor

- Dans la SwissProt
- IL-1F6
- Interleukin-1 epsilon
- IL-1 epsilon
- FIL1 epsilon

UNIPROT

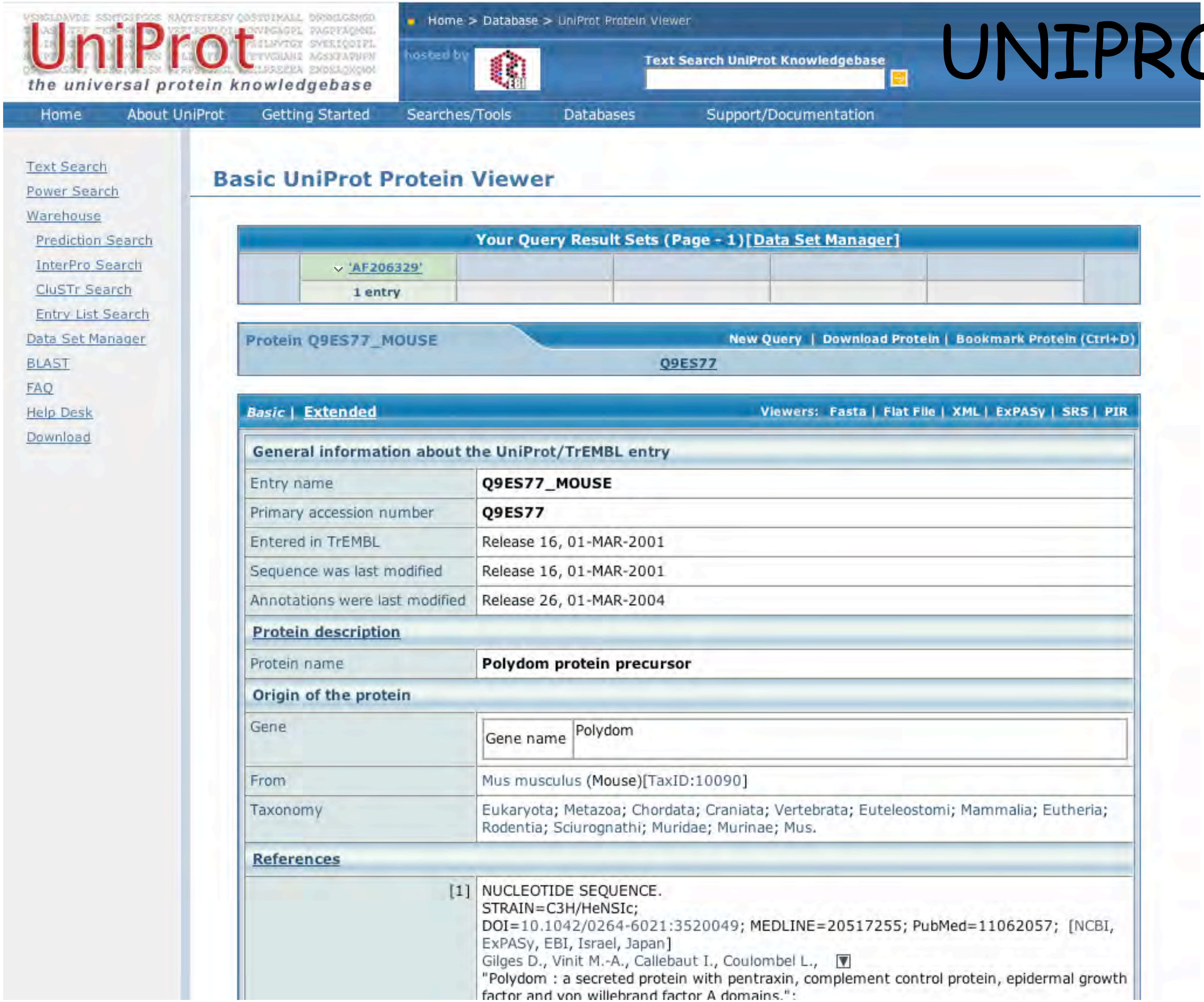
Welcome to UniProt

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components, each optimized for different uses. The **UniProt Knowledgebase (UniProt)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. The **UniProt Non-redundant Reference (UniRef)** databases combine closely related sequences into a single record to speed searches. The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

The sequences and information in UniProt are accessible via **text search**, **BLAST similarity search**, and **FTP**.

UNIPROT



Basic UniProt Protein Viewer

Protein Q9ES77_MOUSE

Q9ES77

General information about the UniProt/TrEMBL entry

Entry name	**Q9ES77_MOUSE**
Primary accession number	**Q9ES77**
Entered in TrEMBL	Release 16, 01-MAR-2001
Sequence was last modified	Release 16, 01-MAR-2001
Annotations were last modified	Release 26, 01-MAR-2004

Protein description

Protein name	**Polydom protein precursor**

Origin of the protein

Gene	Gene name: Polydom
From	Mus musculus (Mouse)[TaxID:10090]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.

References

[1] NUCLEOTIDE SEQUENCE.
STRAIN=C3H/HeNSlc;
DOI=10.1042/0264-6021:3520049; MEDLINE=20517255; PubMed=11062057; [NCBI, ExPASy, EBI, Israel, Japan]
Gilges D., Vinit M.-A., Callebaut I., Coulombel L.,
"Polydom : a secreted protein with pentraxin, complement control protein, epidermal growth factor and von willebrand factor A domains.";

UNIPROT

Polydom

Cross-references	
EMBL	AF206329; AAG32160.1; -.[EMBL/ GenBank/ DDBJ] [CoDingSequence]
HSSP	Q9UCV4; 1NZI.[HSSP/PDB]
Ensembl	ENSMUSG00000028369; Mus musculus.[Entry/Contig]
MGD	MGI:1928849; Polydom.
GO	Molecular function | calcium ion binding | GO:0005509 | inferred from electronic annotation [QuickGO]
InterPro	IPR000152; Asx_hydroxyl_S. IPR000742; EGF_2. IPR001881; EGF_Ca. IPR001438; EGF_II. IPR006209; EGF_like. IPR011641; GCC2_GCC3. IPR003410; Hyalin. IPR001759; Pentaxin. IPR000436; Sushi_SCR_CCP. IPR002035; VWF_A. Graphical view of the domain structure
Pfam	PF00008; EGF; 7. PF07645; EGF_CA; 2. PF07699; GCC2_GCC3; 4. PF02494; HYR; 2. PF00354; Pentaxin; 1. PF00084; Sushi; 33. PF00092; VWA; 1. Pfam graphical view of domain structure
PRINTS	PR00010; EGFBLOOD. PR00895; PENTAXIN. PR00453; VWFADOMAIN.
ProDom	PD002153; Pentaxin; 1. [Domain structure/ List of seq. sharing at least 1 domain]
SMART	SM00032; CCP; 34. SM00179; EGF_CA; 6. SM00327; VWA; 1.
PROSITE	PS00010; ASX_HYDROXYL; 6. PS00022; EGF_1; 9. PS01186; EGF_2; 11. PS50026; EGF_3; 9. PS01187; EGF_CA; 6. PS50825; HYR; 2. PS50923; SUSHI; 34. PS50234; VWFA; 1. PROSITE graphical view of domain structure
Keywords	
	Signal

Motifs

- Motifs consensus (patterns) : Prosite = dictionnaire qui recense les motifs protéiques ayant une signification biologique

- Régions provenant d'un alignement multiple: Block ou PRINTS

PROSITE http://au.expasy.org/prosite/
BLOCKS http://blocks.fhcrc.org/
PRINTS http://umber.sbs.man.ac.uk/dbbrowser/PRINTS

Prosite

Banque de motifs

Motif = signatures protéiques liées à une fonction bio

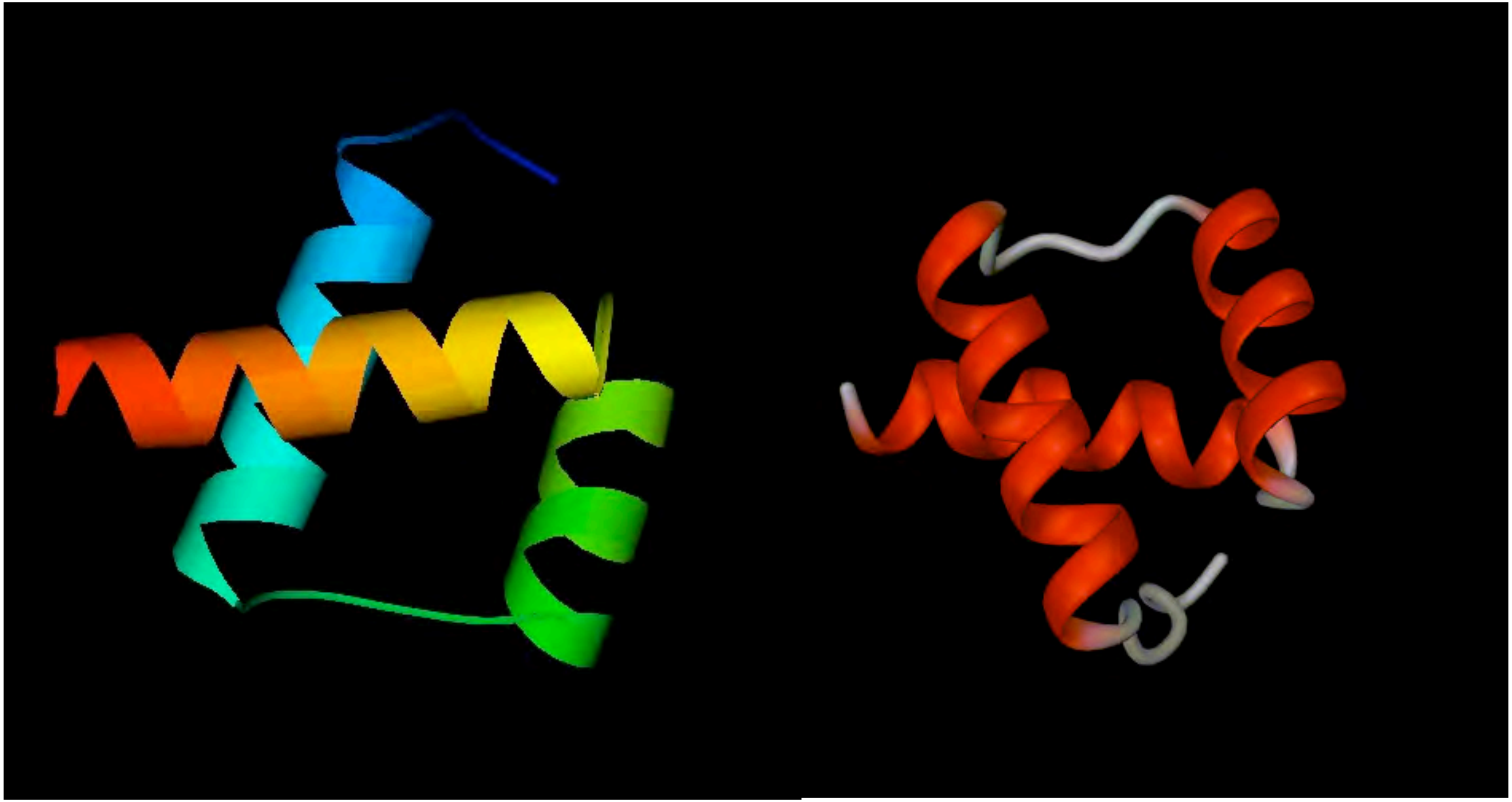
Correspond à un consensus en certaines positions et/ou à une caractéristique structurale permettant de caractériser une famille protéique

Expertise par un réseau de correspondants spécialisés (par fonction)

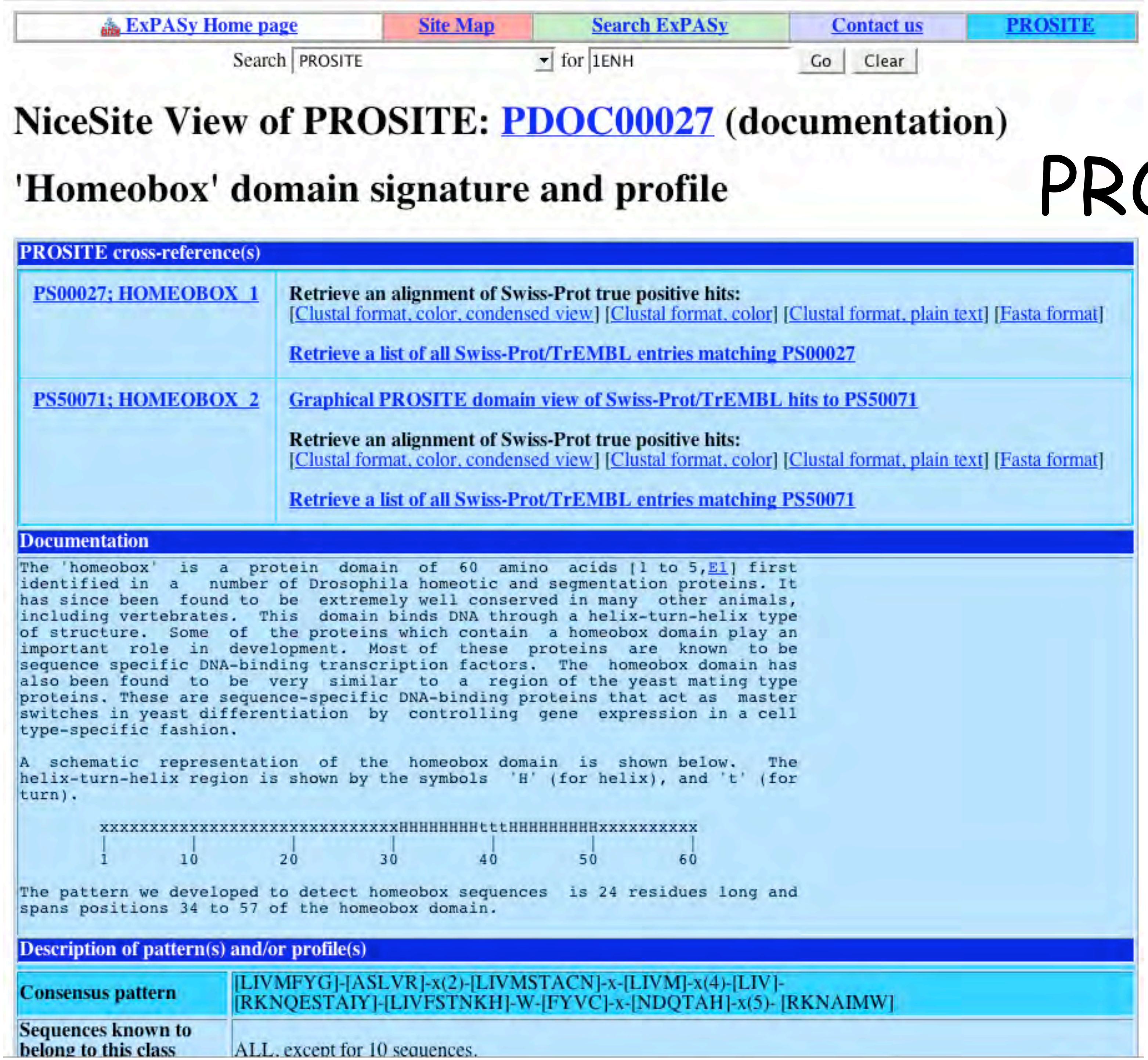
Révision régulière (noms des protéines ont une histoire)

Interrogeable par protéine ou par motif : fournit motif ou famille

1enh Homeodomain



PROSITE



NiceSite View of PROSITE: PDOC00027 (documentation)

'Homeobox' domain signature and profile

PROSITE cross-reference(s)

PS00027; HOMEOBOX_1	**Retrieve an alignment of Swiss-Prot true positive hits:** [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format] Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS00027
PS50071; HOMEOBOX_2	Graphical PROSITE domain view of Swiss-Prot/TrEMBL hits to PS50071 **Retrieve an alignment of Swiss-Prot true positive hits:** [Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format] Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS50071

Documentation

The 'homeobox' is a protein domain of 60 amino acids [1 to 5,E1] first identified in a number of Drosophila homeotic and segmentation proteins. It has since been found to be extremely well conserved in many other animals, including vertebrates. This domain binds DNA through a helix-turn-helix type of structure. Some of the proteins which contain a homeobox domain play an important role in development. Most of these proteins are known to be sequence specific DNA-binding transcription factors. The homeobox domain has also been found to be very similar to a region of the yeast mating type proteins. These are sequence-specific DNA-binding proteins that act as master switches in yeast differentiation by controlling gene expression in a cell type-specific fashion.

A schematic representation of the homeobox domain is shown below. The helix-turn-helix region is shown by the symbols 'H' (for helix), and 't' (for turn).

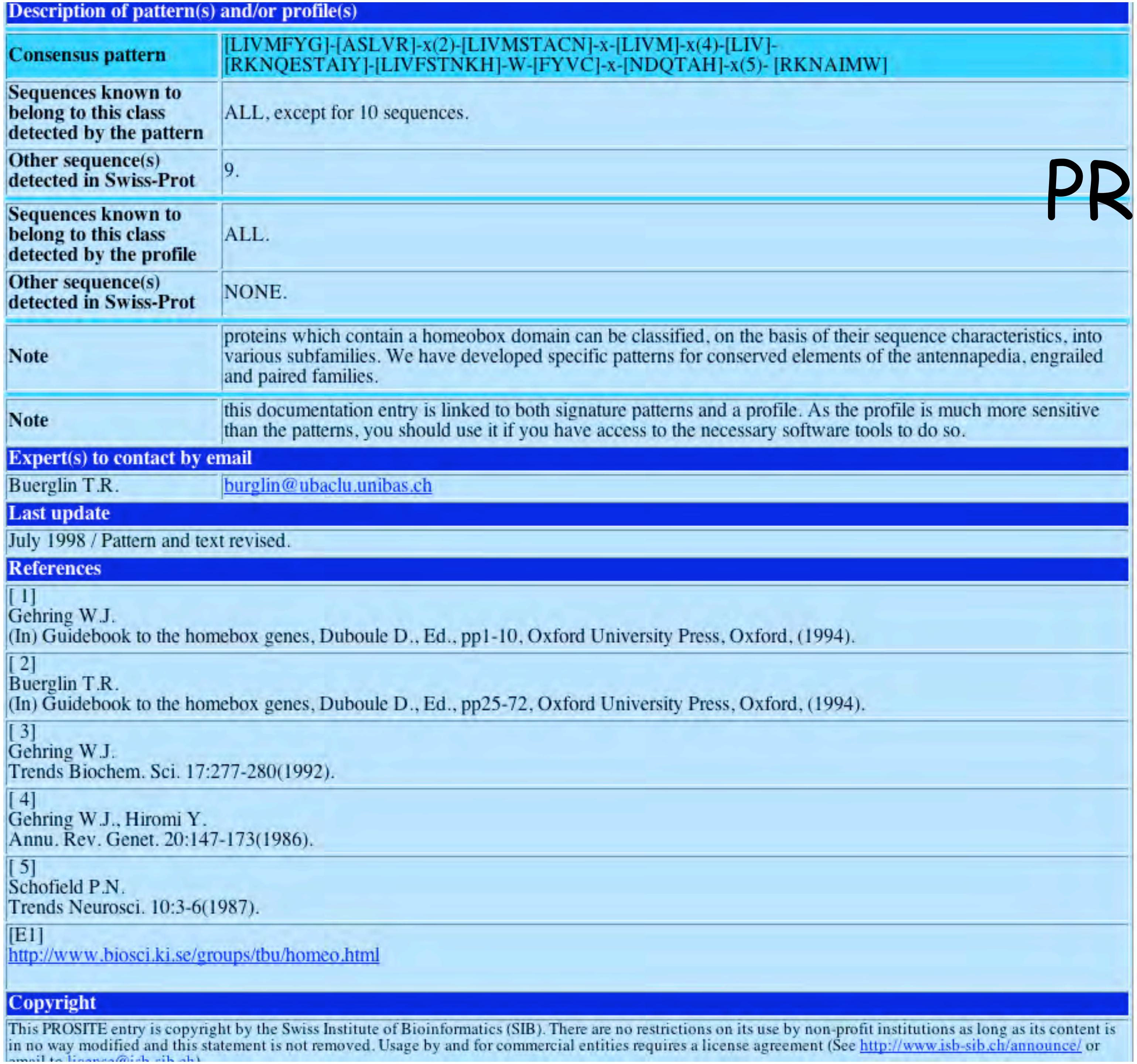
```
        xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxHHHHHHHHtttHHHHHHHHHxxxxxxxxxx
        |        |         |         |         |         |         |
        1        10        20        30        40        50        60
```

The pattern we developed to detect homeobox sequences is 24 residues long and spans positions 34 to 57 of the homeobox domain.

Description of pattern(s) and/or profile(s)

Consensus pattern	[LIVMFYG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)- [RKNAIMW]
Sequences known to belong to this class	ALL, except for 10 sequences.

PROSITE



Description of pattern(s) and/or profile(s)

Consensus pattern	[LIVMFYG]-[ASLVR]-x(2)-[LIVMSTACN]-x-[LIVM]-x(4)-[LIV]-[RKNQESTAIY]-[LIVFSTNKH]-W-[FYVC]-x-[NDQTAH]-x(5)- [RKNAIMW]
Sequences known to belong to this class detected by the pattern	ALL, except for 10 sequences.
Other sequence(s) detected in Swiss-Prot	9.
Sequences known to belong to this class detected by the profile	ALL.
Other sequence(s) detected in Swiss-Prot	NONE.
Note	proteins which contain a homeobox domain can be classified, on the basis of their sequence characteristics, into various subfamilies. We have developed specific patterns for conserved elements of the antennapedia, engrailed and paired families.
Note	this documentation entry is linked to both signature patterns and a profile. As the profile is much more sensitive than the patterns, you should use it if you have access to the necessary software tools to do so.

Expert(s) to contact by email

Buerglin T.R.	burglin@ubaclu.unibas.ch

Last update

July 1998 / Pattern and text revised.

References

[1]
Gehring W.J.
(In) Guidebook to the homebox genes, Duboule D., Ed., pp1-10, Oxford University Press, Oxford, (1994).

[2]
Buerglin T.R.
(In) Guidebook to the homebox genes, Duboule D., Ed., pp25-72, Oxford University Press, Oxford, (1994).

[3]
Gehring W.J.
Trends Biochem. Sci. 17:277-280(1992).

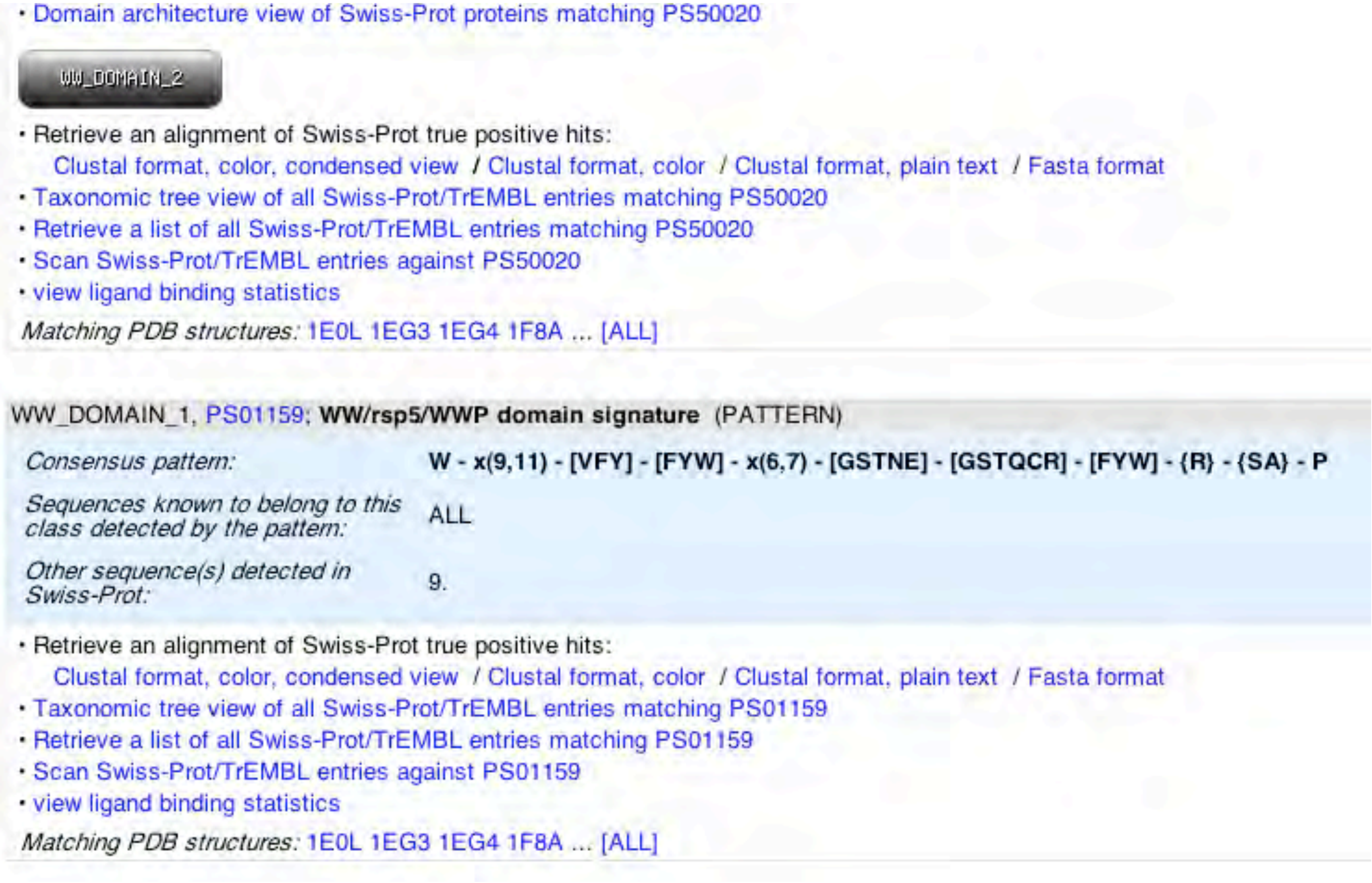
[4]
Gehring W.J., Hiromi Y.
Annu. Rev. Genet. 20:147-173(1986).

[5]
Schofield P.N.
Trends Neurosci. 10:3-6(1987).

[E1]
http://www.biosci.ki.se/groups/tbu/homeo.html

Copyright

PROSITE motif en séquence



- Domain architecture view of Swiss-Prot proteins matching PS50020
- Retrieve an alignment of Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS50020
- Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS50020
- Scan Swiss-Prot/TrEMBL entries against PS50020
- view ligand binding statistics

Matching PDB structures: 1E0L 1EG3 1EG4 1F8A ... [ALL]

WW_DOMAIN_1, PS01159; **WW/rsp5/WWP domain signature** (PATTERN)

Consensus pattern:	**W - x(9,11) - [VFY] - [FYW] - x(6,7) - [GSTNE] - [GSTQCR] - [FYW] - {R} - {SA} - P**
Sequences known to belong to this class detected by the pattern:	ALL
Other sequence(s) detected in Swiss-Prot:	9.

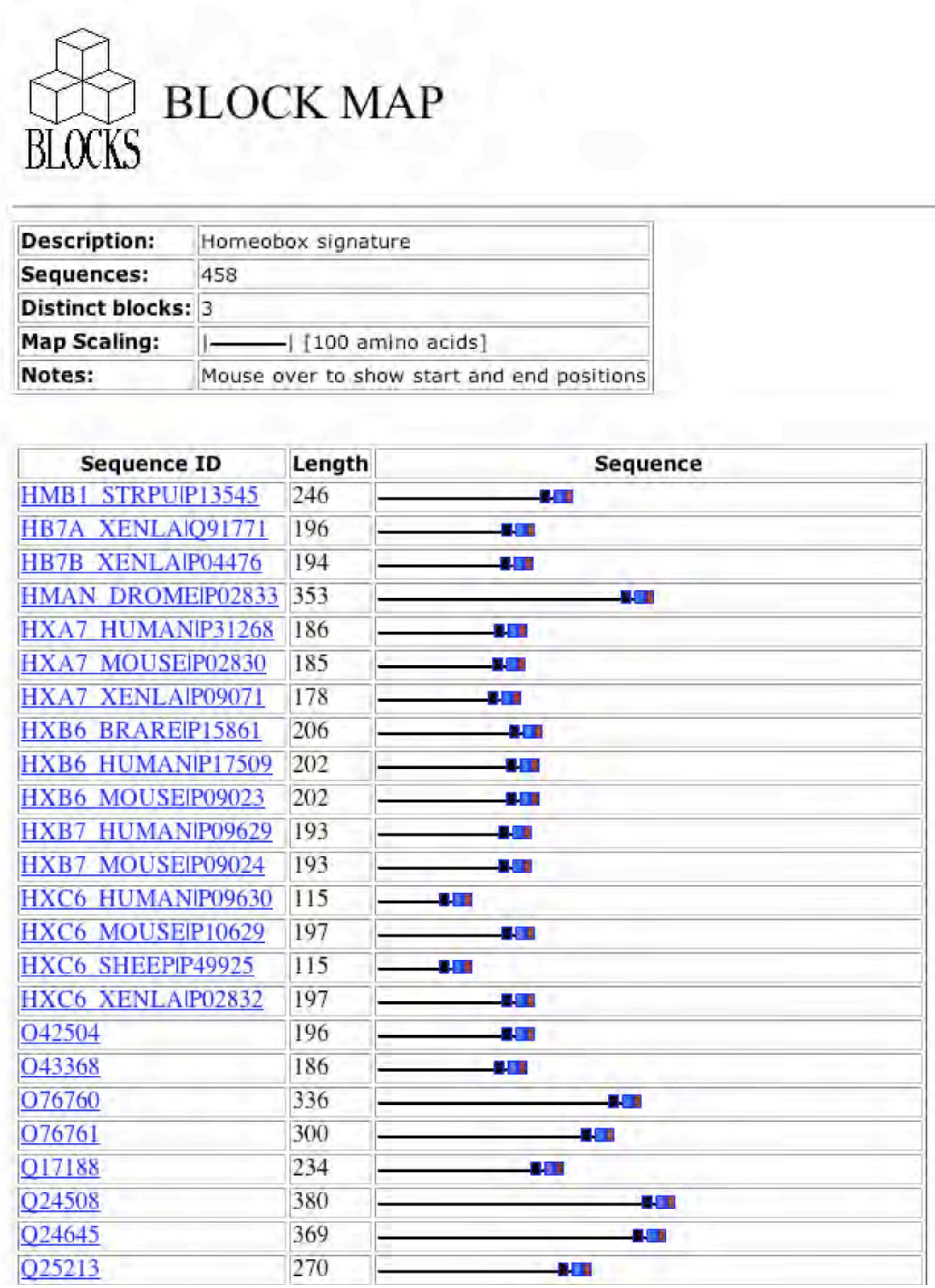
- Retrieve an alignment of Swiss-Prot true positive hits:
  Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Taxonomic tree view of all Swiss-Prot/TrEMBL entries matching PS01159
- Retrieve a list of all Swiss-Prot/TrEMBL entries matching PS01159
- Scan Swiss-Prot/TrEMBL entries against PS01159
- view ligand binding statistics

Matching PDB structures: 1E0L 1EG3 1EG4 1F8A ... [ALL]

Blocks

- Détecte et assemble les régions conservées de protéines apparentées
- Détection par alignement multiples et séparation en blocs de séquences
- Block = superposition de segments protéiques très similaires sans indel
- http://blocks.fhcrc.org/

BLOCKS



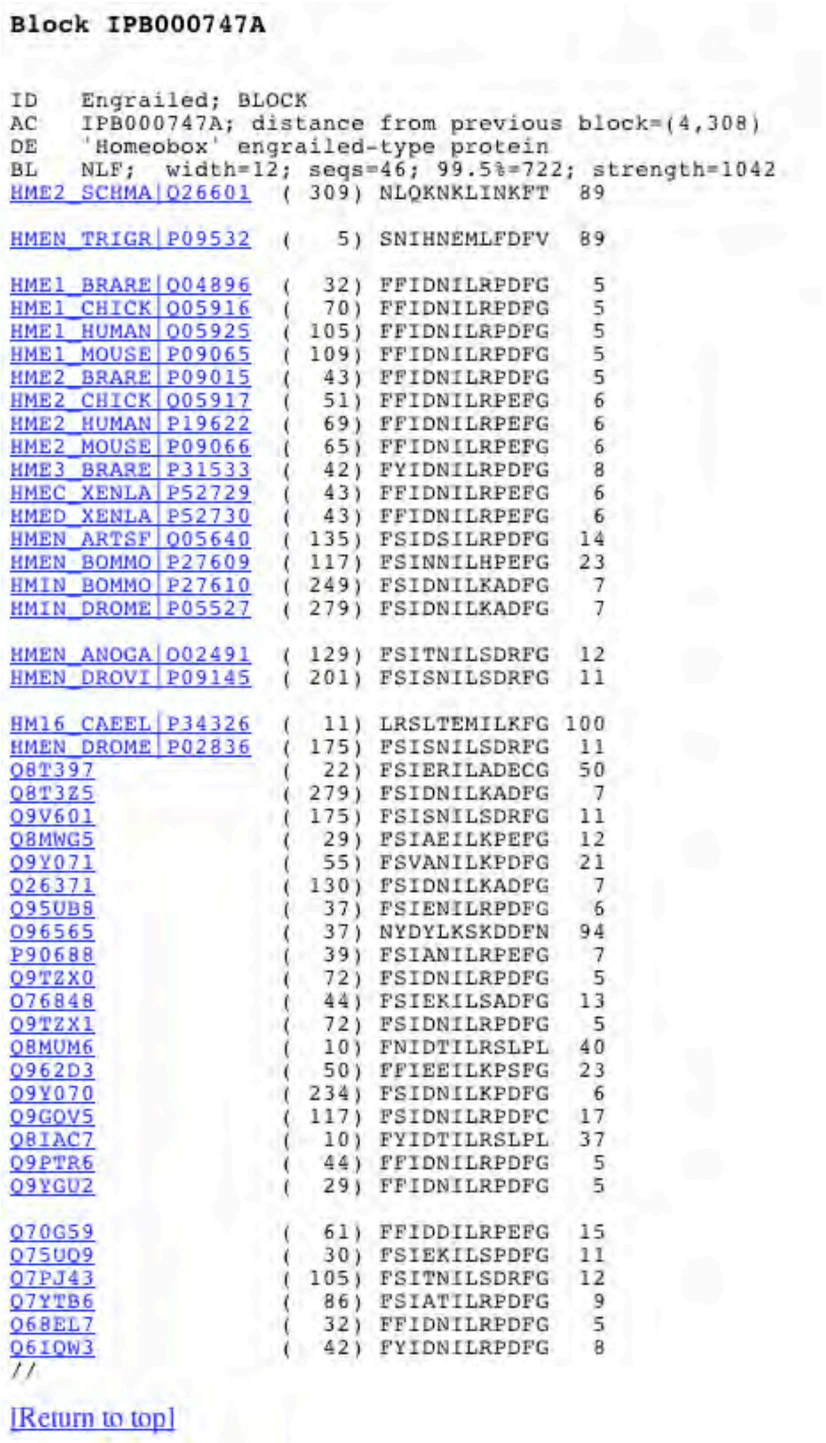
BLOCK MAP

Description:	Homeobox signature
Sequences:	458
Distinct blocks:	3
Map Scaling:	|———| [100 amino acids]
Notes:	Mouse over to show start and end positions

Sequence ID	Length	Sequence
HMB1_STRPU|P13545	246	
HB7A_XENLA|Q91771	196	
HB7B_XENLA|P04476	194	
HMAN_DROME|P02833	353	
HXA7_HUMAN|P31268	186	
HXA7_MOUSE|P02830	185	
HXA7_XENLA|P09071	178	
HXB6_BRARE|P15861	206	
HXB6_HUMAN|P17509	202	
HXB6_MOUSE|P09023	202	
HXB7_HUMAN|P09629	193	
HXB7_MOUSE|P09024	193	
HXC6_HUMAN|P09630	115	
HXC6_MOUSE|P10629	197	
HXC6_SHEEP|P49925	115	
HXC6_XENLA|P02832	197	
O42504	196	
O43368	186	
O76760	336	
O76761	300	
Q17188	234	
Q24508	380	
Q24645	369	
Q25213	270	

1enh

BLOCKS



Block IPB000747A

```
ID   Engrailed; BLOCK
AC   IPB000747A; distance from previous block=(4,308)
DE   'Homeobox' engrailed-type protein
BL   NLF;  width=12; seqs=46; 99.5%=722; strength=1042
HME2_SCHMA|Q26601  ( 309) NLQKNKLINKFT  89

HMEN_TRIGR|P09532  (   5) SNIHNEMLFDFV  89

HME1_BRARE|Q04896  (  32) FFIDNILRPDFG   5
HME1_CHICK|Q05916  (  70) FFIDNILRPDFG   5
HME1_HUMAN|Q05925  ( 105) FFIDNILRPDFG   5
HME1_MOUSE|P09065  ( 109) FFIDNILRPDFG   5
HME2_BRARE|P09015  (  43) FFIDNILRPDFG   5
HME2_CHICK|Q05917  (  51) FFIDNILRPEFG   6
HME2_HUMAN|P19622  (  69) FFIDNILRPEFG   6
HME2_MOUSE|P09066  (  65) FFIDNILRPEFG   6
HME3_BRARE|P31533  (  42) FYIDNILRPDFG   8
HMEC_XENLA|P52729  (  43) FFIDNILRPEFG   6
HMED_XENLA|P52730  (  43) FFIDNILRPEFG   6
HMEN_ARTSF|Q05640  ( 135) FSIDSILRPDFG  14
HMEN_BOMMO|P27609  ( 117) FSINNILHPEFG  23
HMIN_BOMMO|P27610  ( 249) FSIDNILKADFG   7
HMIN_DROME|P05527  ( 279) FSIDNILKADFG   7

HMEN_ANOGA|O02491  ( 129) FSITNILSDRFG  12
HMEN_DROVI|P09145  ( 201) FSISNILSDRFG  11

HM16_CAEEL|P34326  (  11) LRSLTEMILKFG 100
HMEN_DROME|P02836  ( 175) FSISNILSDRFG  11
Q8T397             (  22) FSIERILADECG  50
Q8T3Z5             ( 279) FSIDNILKADFG   7
Q9V601             ( 175) FSISNILSDRFG  11
Q8MWG5             (  29) FSIAEILKPEFG  12
Q9Y071             (  55) FSVANILKPDFG  21
Q26371             ( 130) FSIDNILKADFG   7
Q95UB8             (  37) FSIENILRPDFG   6
O96565             (  37) NYDYLKSKDDFN  94
P90688             (  39) FSIANILRPEFG   7
Q9TZX0             (  72) FSIDNILRPDFG   5
O76848             (  44) FSIEKILSADFG  13
Q9TZX1             (  72) FSIDNILRPDFG   5
Q8MUM6             (  10) FNIDTILRSLPL  40
Q962D3             (  50) FFIEEILKPSFG  23
Q9Y070             ( 234) FSIDNILKPDFG   6
Q9GOV5             ( 117) FSIDNILRPDFC  17
Q8IAC7             (  10) FYIDTILRSLPL  37
Q9PTR6             (  44) FFIDNILRPDFG   5
Q9YGU2             (  29) FFIDNILRPDFG   5

Q70G59             (  61) FFIDDILRPEFG  15
Q75UQ9             (  30) FSIEKILSPDFG  11
Q7PJ43             ( 105) FSITNILSDRFG  12
Q7YTB6             (  86) FSIATILRPDFG   9
Q68EL7             (  32) FFIDNILRPDFG   5
Q6IQW3             (  42) FYIDNILRPDFG   8
//
```

[Return to top]

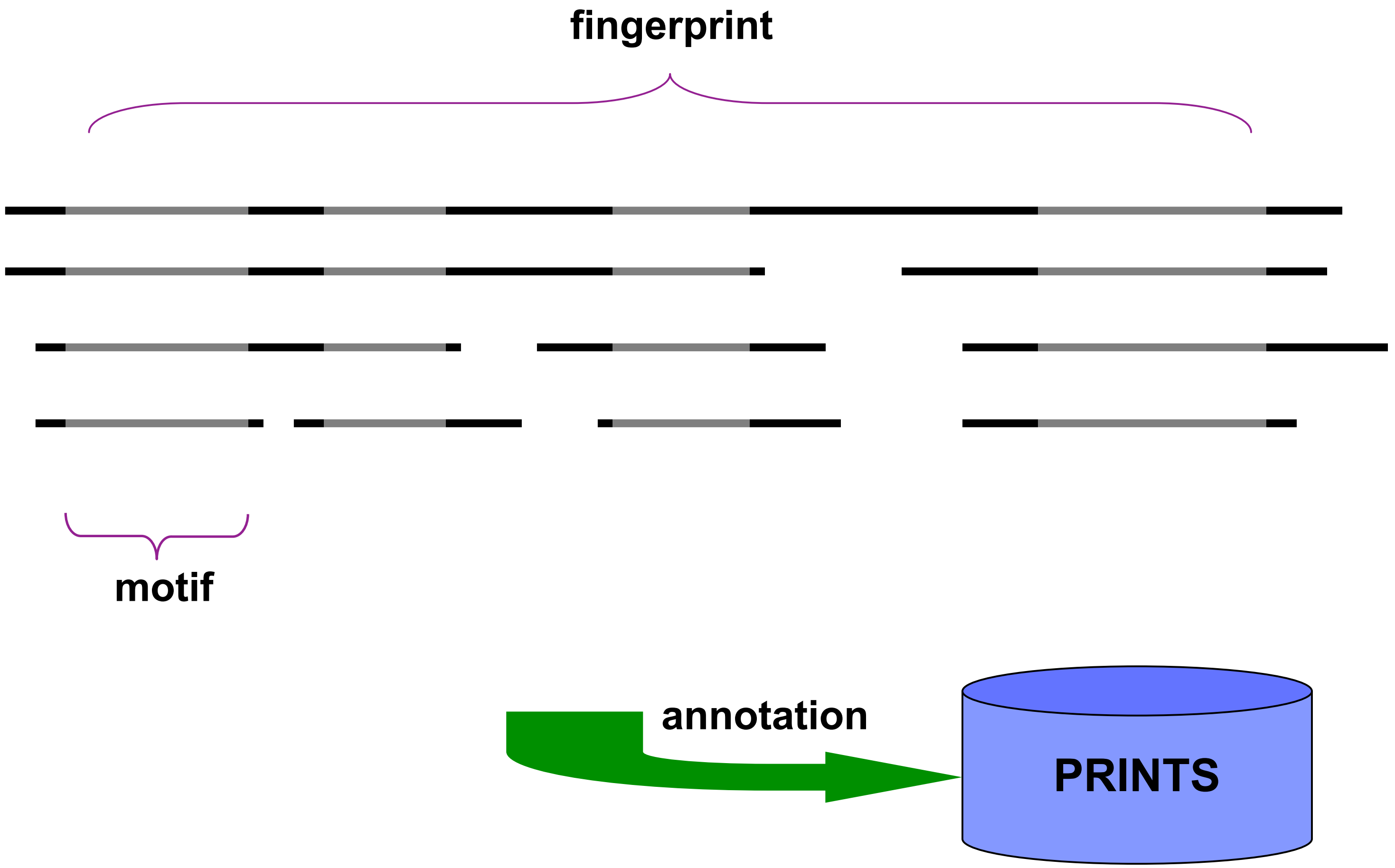
1enh

PRINTS

- Fingerprints (Empreintes)– groupes de motifs extraits d’alignements multiples de séquences
- Signatures d’une famille protéique
- PRINTS-BLAST : Permet de rechercher un motif de la banque PRINTS avec le logiciel BLAST

PRINTS

What is a fingerprint?



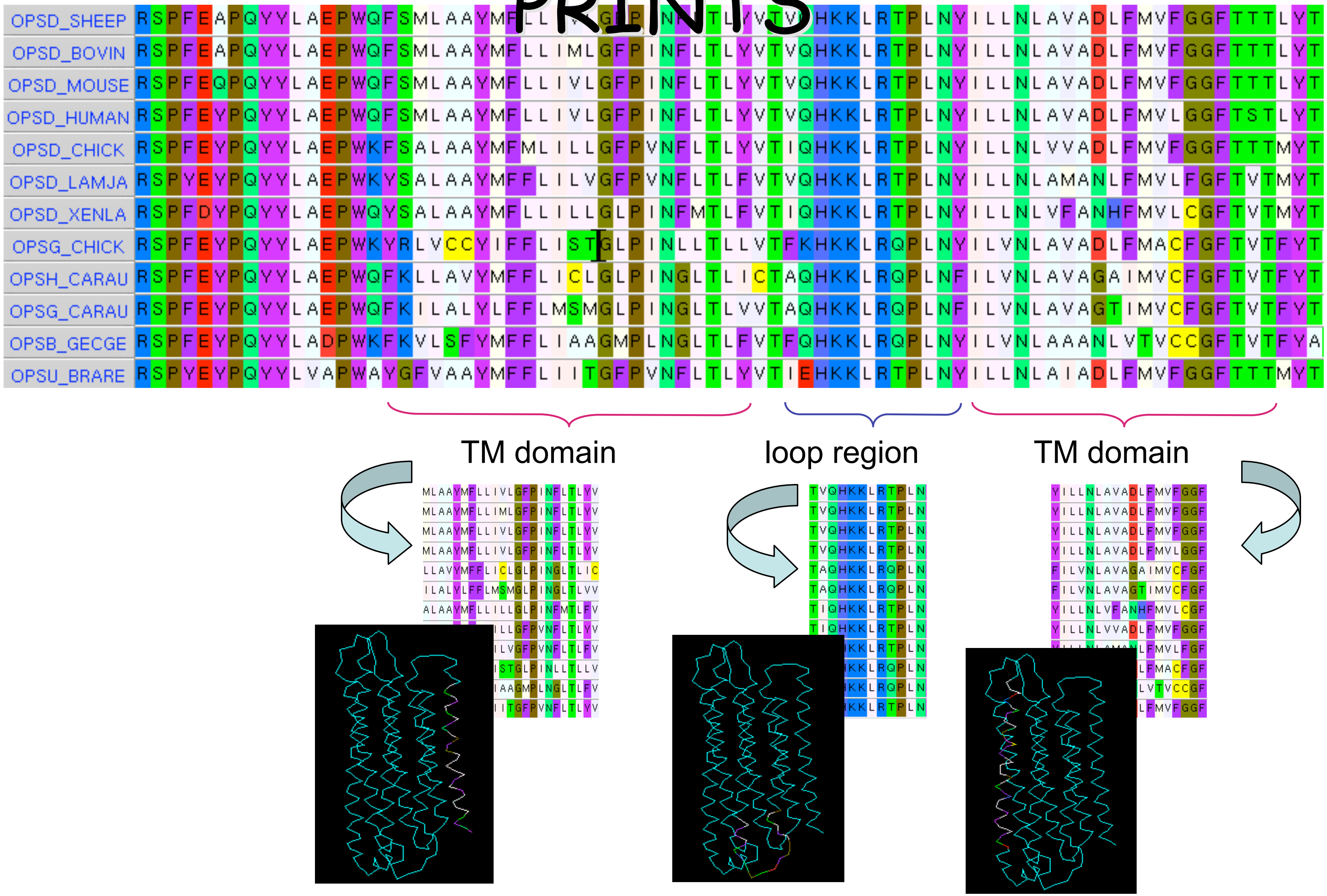
fingerprint

motif

annotation

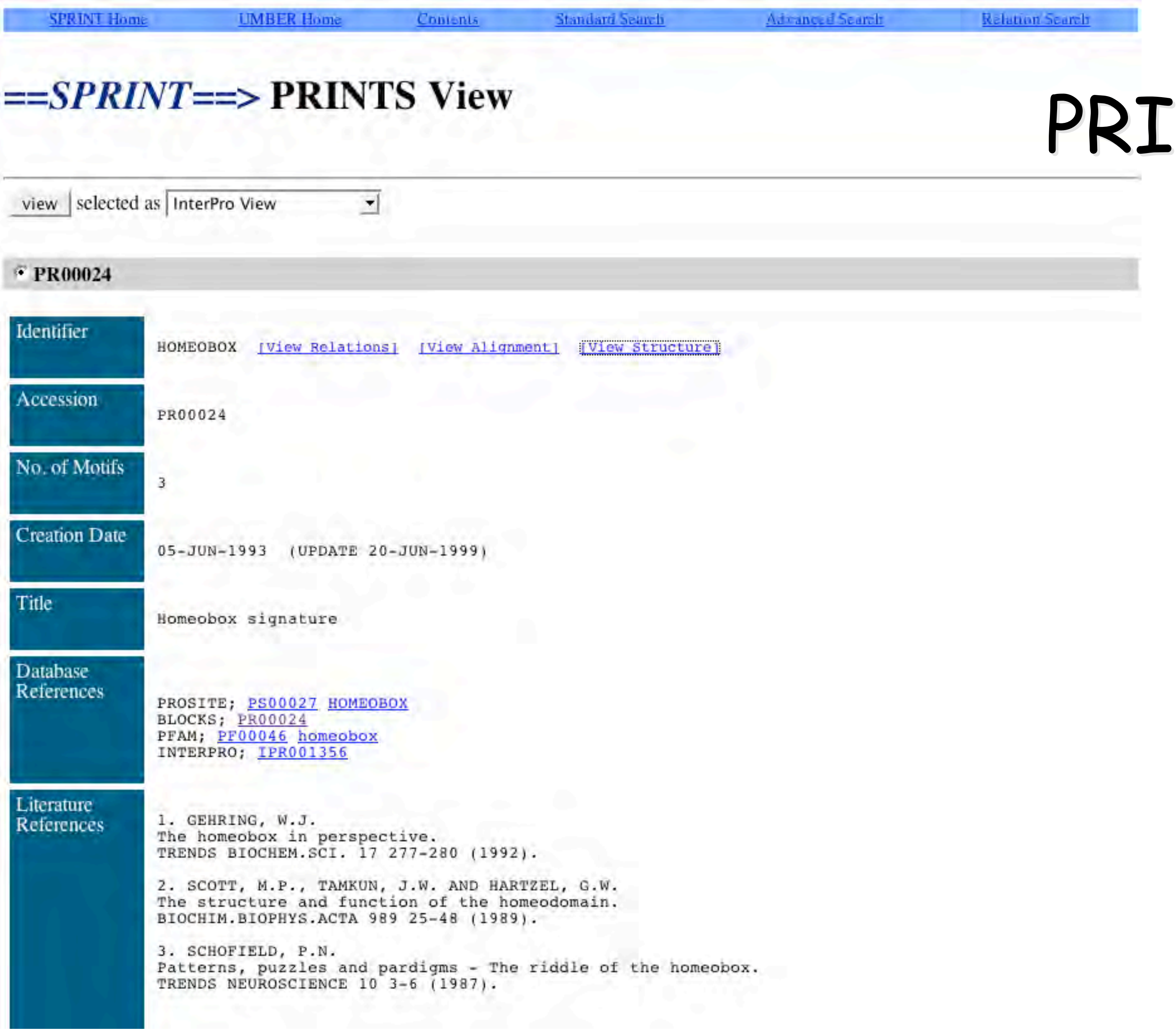
PRINTS

PRINTS



```
OPSD_SHEEP RSPFEAPQYYLAEPWQFSMLAAYMFLL[illegible]V[illegible]GFP[illegible]NF[illegible]TLYVTV[illegible]HKKLRTPLNYILLNLAVADLFMVFGGFTTTLYT
OPSD_BOVIN RSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYT
OPSD_MOUSE RSPFEQPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYT
OPSD_HUMAN RSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYT
OPSD_CHICK RSPFEYPQYYLAEPWKFSALAAYMFMLILLGFPVNFLTLYVTIQHKKLRTPLNYILLNLVVADLFMVFGGFTTTMYT
OPSD_LAMJA RSPYEYPQYYLAEPWKYSALAAYMFFLILVGFPVNFLTLFVTVQHKKLRTPLNYILLNLAMANLFMVLFGFTVTMYT
OPSD_XENLA RSPFDYPQYYLAEPWQYSALAAYMFLLILLGLPINFMTLFVTIQHKKLRTPLNYILLNLVFANHFMVLCGFTVTMYT
OPSG_CHICK RSPFEYPQYYLAEPWKYRLVCCYIFFLISTGLPINLLTLLVTFKHKKLRQPLNYILVNLAVADLFMACFGFTVTFYT
OPSH_CARAU RSPFEYPQYYLAEPWQFKLLAVYMFFLICLGLPINGLTLICTAQHKKLRQPLNFILVNLAVAGAIMVCFGFTVTFYT
OPSG_CARAU RSPFEYPQYYLAEPWQFKILALYLFFLMSMGLPINGLTLVVTAQHKKLRQPLNFILVNLAVAGTIMVCFGFTVTFYT
OPSB_GECGE RSPFEYPQYYLADPWKFKVLSFYMFFLIAAGMPLNGLTLFVTFQHKKLRQPLNYILVNLAAANLVTVCCGFTVTFYA
OPSU_BRARE RSPYEYPQYYLVAPWAYGFVAAYMFFLIITGFPVNFLTLYVTIEHKKLRTPLNYILLNLAIADLFMVFGGFTTTMYT
```

PRINTS



==*SPRINT*==> **PRINTS View**

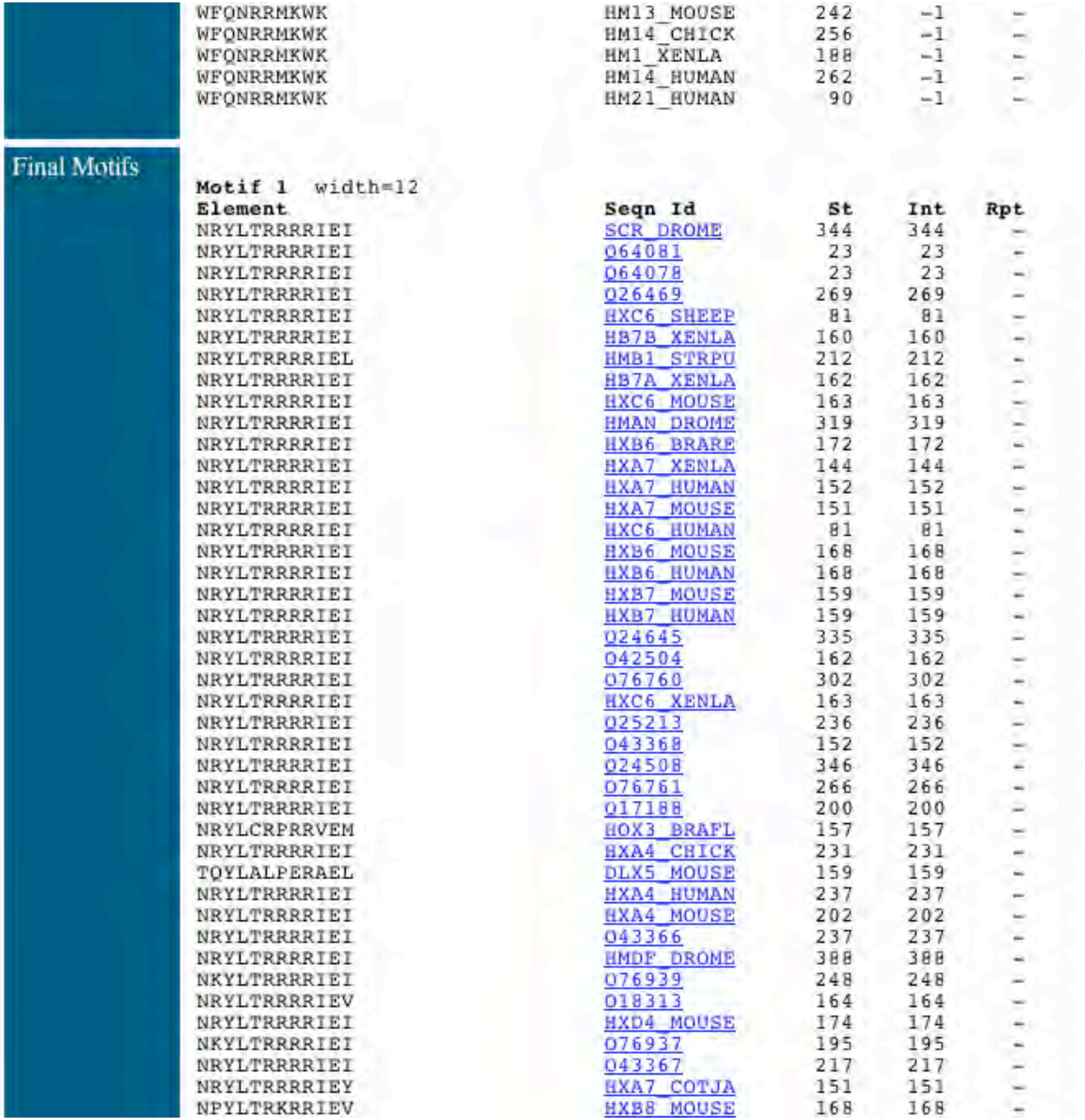
PR00024

Identifier	HOMEOBOX [View Relations] [View Alignment] [View Structure]
Accession	PR00024
No. of Motifs	3
Creation Date	05-JUN-1993 (UPDATE 20-JUN-1999)
Title	Homeobox signature
Database References	PROSITE; PS00027 HOMEOBOX BLOCKS; PR00024 PFAM; PF00046 homeobox INTERPRO; IPR001356
Literature References	1. GEHRING, W.J. The homeobox in perspective. TRENDS BIOCHEM.SCI. 17 277-280 (1992). 2. SCOTT, M.P., TAMKUN, J.W. AND HARTZEL, G.W. The structure and function of the homeodomain. BIOCHIM.BIOPHYS.ACTA 989 25-48 (1989). 3. SCHOFIELD, P.N. Patterns, puzzles and pardigms - The riddle of the homeobox. TRENDS NEUROSCIENCE 10 3-6 (1987).

1enh

PRINTS

1enh



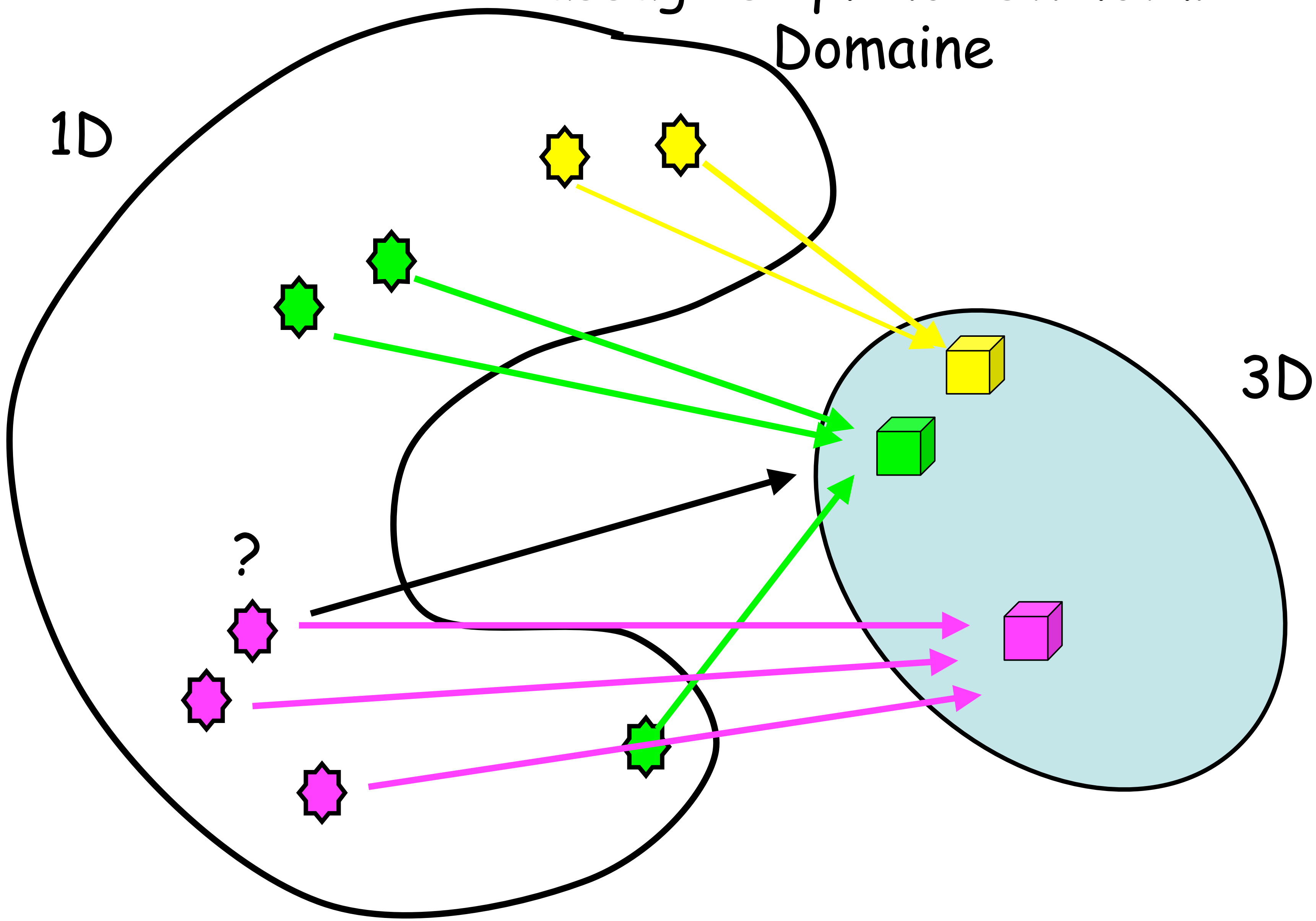
```
WFQNRRMKWK                    HM13_MOUSE      242     -1      -
WFQNRRMKWK                    HM14_CHICK      256     -1      -
WFQNRRMKWK                    HM1_XENLA       188     -1      -
WFQNRRMKWK                    HM14_HUMAN      262     -1      -
WFQNRRMKWK                    HM21_HUMAN       90     -1      -

Motif 1  width=12
Element                       Seqn Id          St    Int    Rpt
NRYLTRRRRIEI                  SCR_DROME       344    344      -
NRYLTRRRRIEI                  Q64081           23     23      -
NRYLTRRRRIEI                  Q64078           23     23      -
NRYLTRRRRIEI                  Q26469          269    269      -
NRYLTRRRRIEI                  HXC6_SHEEP       81     81      -
NRYLTRRRRIEI                  HB7B_XENLA      160    160      -
NRYLTRRRRIEL                  HMB1_STRPU      212    212      -
NRYLTRRRRIEI                  HB7A_XENLA      162    162      -
NRYLTRRRRIEI                  HXC6_MOUSE      163    163      -
NRYLTRRRRIEI                  HMAN_DROME      319    319      -
NRYLTRRRRIEI                  HXB6_BRARE      172    172      -
NRYLTRRRRIEI                  HXA7_XENLA      144    144      -
NRYLTRRRRIEI                  HXA7_HUMAN      152    152      -
NRYLTRRRRIEI                  HXA7_MOUSE      151    151      -
NRYLTRRRRIEI                  HXC6_HUMAN       81     81      -
NRYLTRRRRIEI                  HXB6_MOUSE      168    168      -
NRYLTRRRRIEI                  HXB6_HUMAN      168    168      -
NRYLTRRRRIEI                  HXB7_MOUSE      159    159      -
NRYLTRRRRIEI                  HXB7_HUMAN      159    159      -
NRYLTRRRRIEI                  Q24645          335    335      -
NRYLTRRRRIEI                  O42504          162    162      -
NRYLTRRRRIEI                  O76760          302    302      -
NRYLTRRRRIEI                  HXC6_XENLA      163    163      -
NRYLTRRRRIEI                  Q25213          236    236      -
NRYLTRRRRIEI                  O43368          152    152      -
NRYLTRRRRIEI                  Q24508          346    346      -
NRYLTRRRRIEI                  O76761          266    266      -
NRYLTRRRRIEI                  Q17188          200    200      -
NRYLCRPRRVEM                  HOX3_BRAFL      157    157      -
NRYLTRRRRIEI                  HXA4_CHICK      231    231      -
TQYLALPERAEL                  DLX5_MOUSE      159    159      -
NRYLTRRRRIEI                  HXA4_HUMAN      237    237      -
NRYLTRRRRIEI                  HXA4_MOUSE      202    202      -
NRYLTRRRRIEI                  O43366          237    237      -
NRYLTRRRRIEI                  HMDF_DROME      388    388      -
NKYLTRRRRIEI                  O76939          248    248      -
NRYLTRRRRIEV                  O18313          164    164      -
NRYLTRRRRIEI                  HXD4_MOUSE      174    174      -
NKYLTRRRRIEI                  O76937          195    195      -
NRYLTRRRRIEI                  O43367          217    217      -
NRYLTRRRRIEY                  HXA7_COTJA      151    151      -
NPYLTRKRRIEV                  HXB8_MOUSE      168    168      -
```

Séquence et Structure ne sont pas sur un même bateau

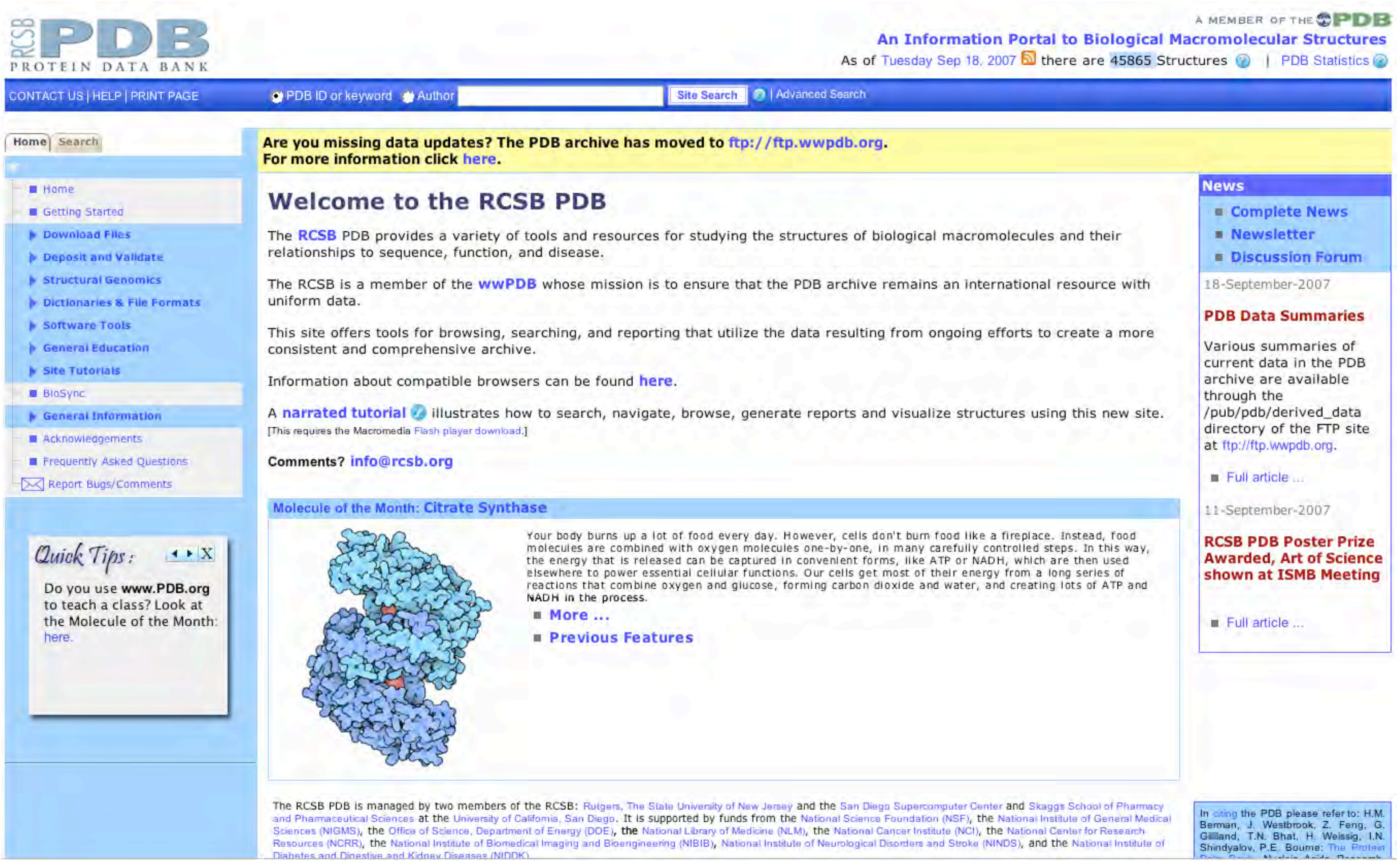
- Biosphère = 10^n séquences, avec n voisin de 12
- Nombre de repliements (architectures) différents : 10^3
- Deux séquences voisines se replient en deux structures voisines
- Deux structures voisines peuvent porter des séquences très différentes

Passage séquence structure
Domaine



Banques 3D

Banque de structures



An Information Portal to Biological Macromolecular Structures

As of Tuesday Sep 18, 2007 there are 45865 Structures | PDB Statistics

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.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site.
[This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Citrate Synthase

Your body burns up a lot of food every day. However, cells don't burn food like a fireplace. Instead, food molecules are combined with oxygen molecules one-by-one, in many carefully controlled steps. In this way, the energy that is released can be captured in convenient forms, like ATP or NADH, which are then used elsewhere to power essential cellular functions. Our cells get most of their energy from a long series of reactions that combine oxygen and glucose, forming carbon dioxide and water, and creating lots of ATP and NADH in the process.

- **More ...**
- **Previous Features**

Do you use **www.PDB.org** to teach a class? Look at the Molecule of the Month: here.

18-September-2007

PDB Data Summaries

Various summaries of current data in the PDB archive are available through the /pub/pdb/derived_data directory of the FTP site at ftp://ftp.wwpdb.org.

- Full article ...

11-September-2007

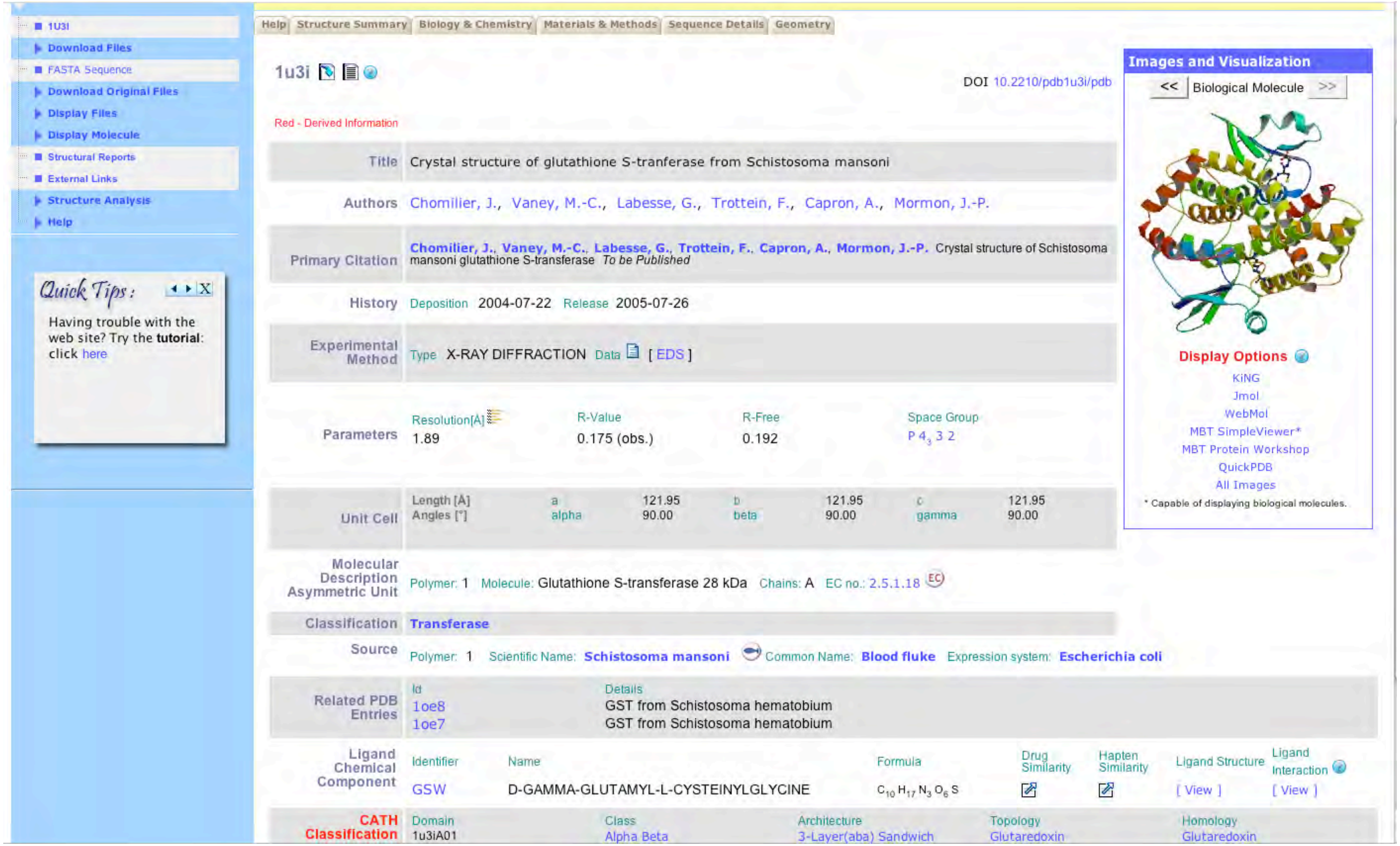
RCSB PDB Poster Prize Awarded, Art of Science shown at ISMB Meeting

- Full article ...

The RCSB PDB is managed by two members of the RCSB: Rutgers, The State University of New Jersey and the San Diego Supercomputer Center and Skaggs School of Pharmacy and Pharmaceutical Sciences at the University of California, San Diego. It is supported by funds from the National Science Foundation (NSF), the National Institute of General Medical Sciences (NIGMS), the Office of Science, Department of Energy (DOE), the National Library of Medicine (NLM), the National Cancer Institute (NCI), the National Center for Research Resources (NCRR), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institute of Neurological Disorders and Stroke (NINDS), and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

In citing the PDB please refer to: H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein

PDB



Quick Tips:

Having trouble with the web site? Try the **tutorial**: click here

1u3i

DOI 10.2210/pdb1u3i/pdb

Red - Derived Information

Title	Crystal structure of glutathione S-tranferase from Schistosoma mansoni
Authors	Chomilier, J., Vaney, M.-C., Labesse, G., Trottein, F., Capron, A., Mormon, J.-P.
Primary Citation	**Chomilier, J., Vaney, M.-C., Labesse, G., Trottein, F., Capron, A., Mormon, J.-P.** Crystal structure of Schistosoma mansoni glutathione S-transferase *To be Published*
History	Deposition 2004-07-22 Release 2005-07-26
Experimental Method	Type X-RAY DIFFRACTION Data [EDS]

Parameters

Resolution[Å]	R-Value	R-Free	Space Group
1.89	0.175 (obs.)	0.192	$P\ 4_3\ 3\ 2$

Unit Cell

Length [Å]	a	121.95	b	121.95	c	121.95
Angles [°]	alpha	90.00	beta	90.00	gamma	90.00

Molecular Description Asymmetric Unit Polymer: 1 Molecule: Glutathione S-transferase 28 kDa Chains: A EC no.: 2.5.1.18

Classification **Transferase**

Source Polymer: 1 Scientific Name: **Schistosoma mansoni** Common Name: **Blood fluke** Expression system: **Escherichia coli**

Related PDB Entries

Id	Details
1oe8	GST from Schistosoma hematobium
1oe7	GST from Schistosoma hematobium

Ligand Chemical Component

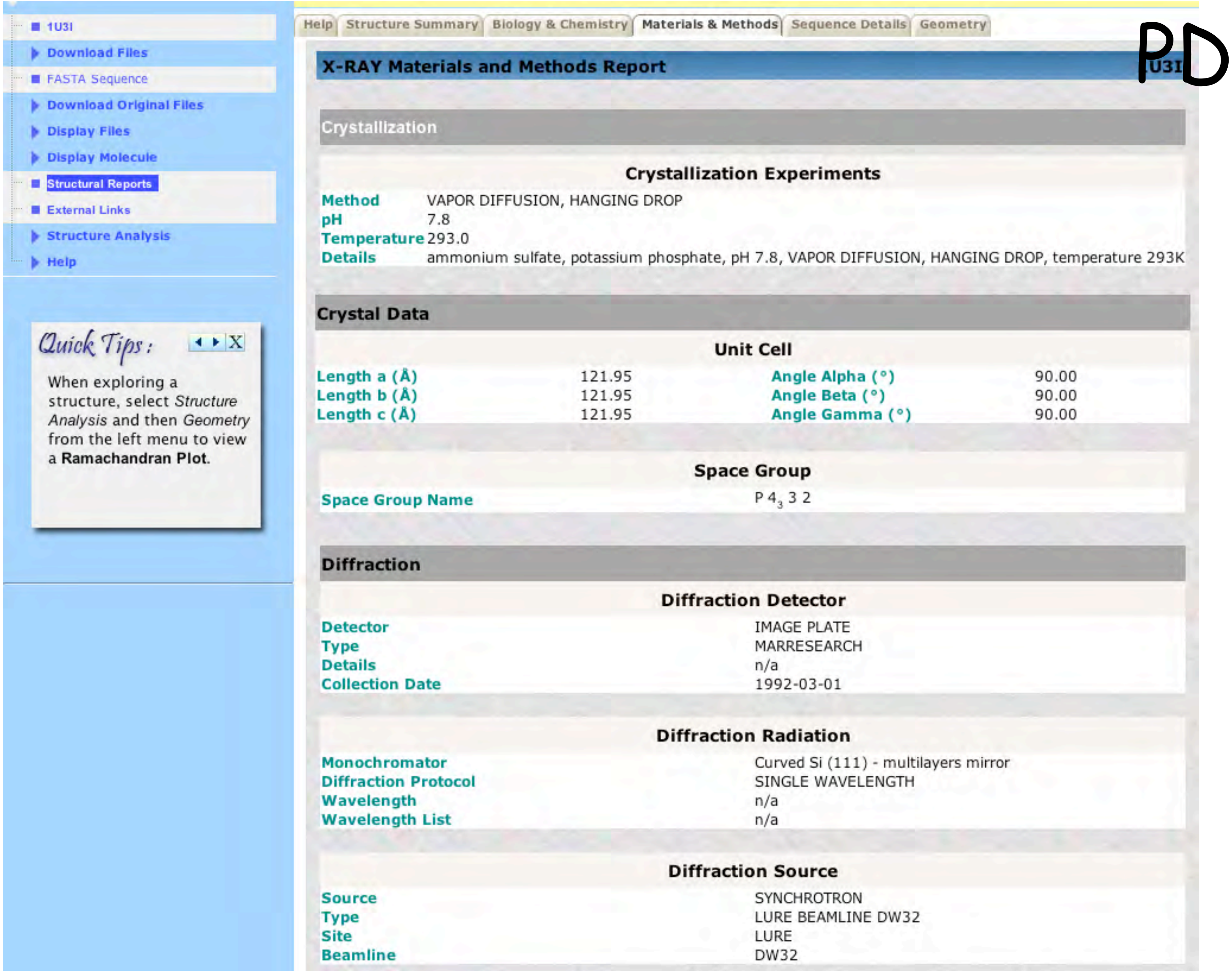
Identifier	Name	Formula	Drug Similarity	Hapten Similarity	Ligand Structure	Ligand Interaction
GSW	D-GAMMA-GLUTAMYL-L-CYSTEINYLGLYCINE	$C_{10}H_{17}N_3O_6S$			[View]	[View]

CATH Classification

Domain	Class	Architecture	Topology	Homology
1u3iA01	Alpha Beta	3-Layer(aba) Sandwich	Glutaredoxin	Glutaredoxin

* Capable of displaying biological molecules.

PDB



When exploring a structure, select *Structure Analysis* and then *Geometry* from the left menu to view a **Ramachandran Plot**.

X-RAY Materials and Methods Report

Crystallization

Crystallization Experiments

Method	VAPOR DIFFUSION, HANGING DROP
pH	7.8
Temperature	293.0
Details	ammonium sulfate, potassium phosphate, pH 7.8, VAPOR DIFFUSION, HANGING DROP, temperature 293K

Crystal Data

Unit Cell

Length a (Å)	121.95	**Angle Alpha (°)**	90.00
Length b (Å)	121.95	**Angle Beta (°)**	90.00
Length c (Å)	121.95	**Angle Gamma (°)**	90.00

Space Group

Space Group Name	P 4_3 3 2

Diffraction

Diffraction Detector

Detector	IMAGE PLATE
Type	MARRESEARCH
Details	n/a
Collection Date	1992-03-01

Diffraction Radiation

Monochromator	Curved Si (111) - multilayers mirror
Diffraction Protocol	SINGLE WAVELENGTH
Wavelength	n/a
Wavelength List	n/a

Diffraction Source

Source	SYNCHROTRON
Type	LURE BEAMLINE DW32
Site	LURE
Beamline	DW32

PDB

Sequence Details

Chain A (polymer 1)

Description	Glutathione S-transferase 28 kDa
Chain Type	polypeptide(L)
UniProt reference	P09792 [show this sequence below]
Length	211 residues
CATH domain assignment [hide] [reference]	1u3iA01 80 residues 1u3iA02 128 residues
DSSP secondary structure [hide] [reference]	54% helical (11 helices; 116 residues) 10% beta sheet (4 strands; 22 residues)
More annotations	

Quick Tips:

When exploring a structure, select *Structure Analysis* and then *Geometry* from the left menu to view a **Ramachandran Plot**.

MAGEHIKVIYFDGRGRAESIRMTLVAAGVDYEDERISFQDWPKIKPTIPGGRLPAVKVTD

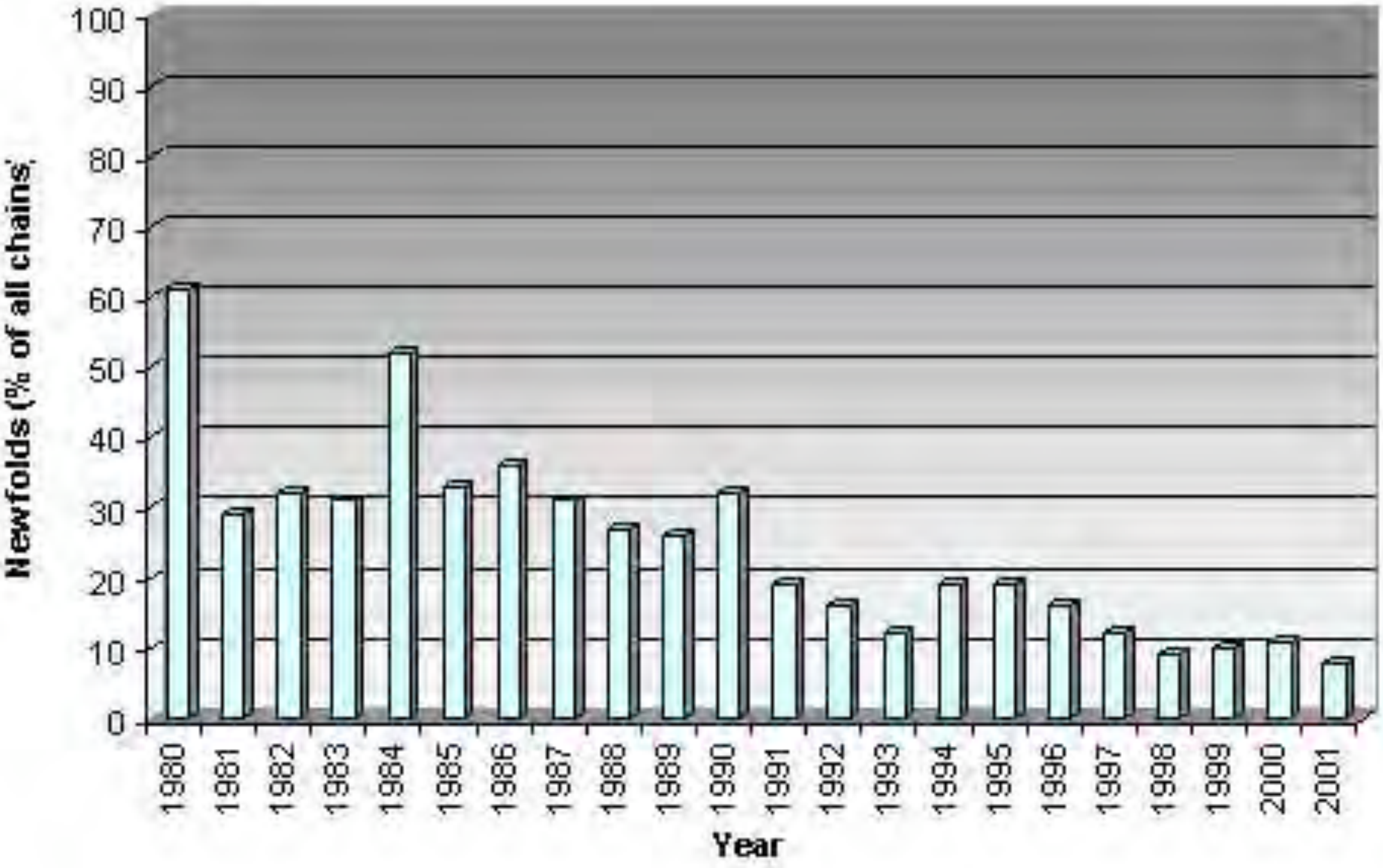
DHGHVKWMLESLAIARYMAKKHHMMGETDEEYYSVEKLIGQAEDVEHEYHKTLMKPQEEK

EKITKEILNGKVPVLFNMICESLKGSTGKLAVGDKVTLADLVLIAVIDHVTDLDKGFLTG

KYPEIHKHRENLLASSPRLAKYLSNRPATPF

PDB

Taux de nouveaux repliements décrits annuellement



Fichier PDB

Coordonnées atomiques disponibles dans une banque de structures, la Protein Data Bank http://www.rcsb.org/pdb/

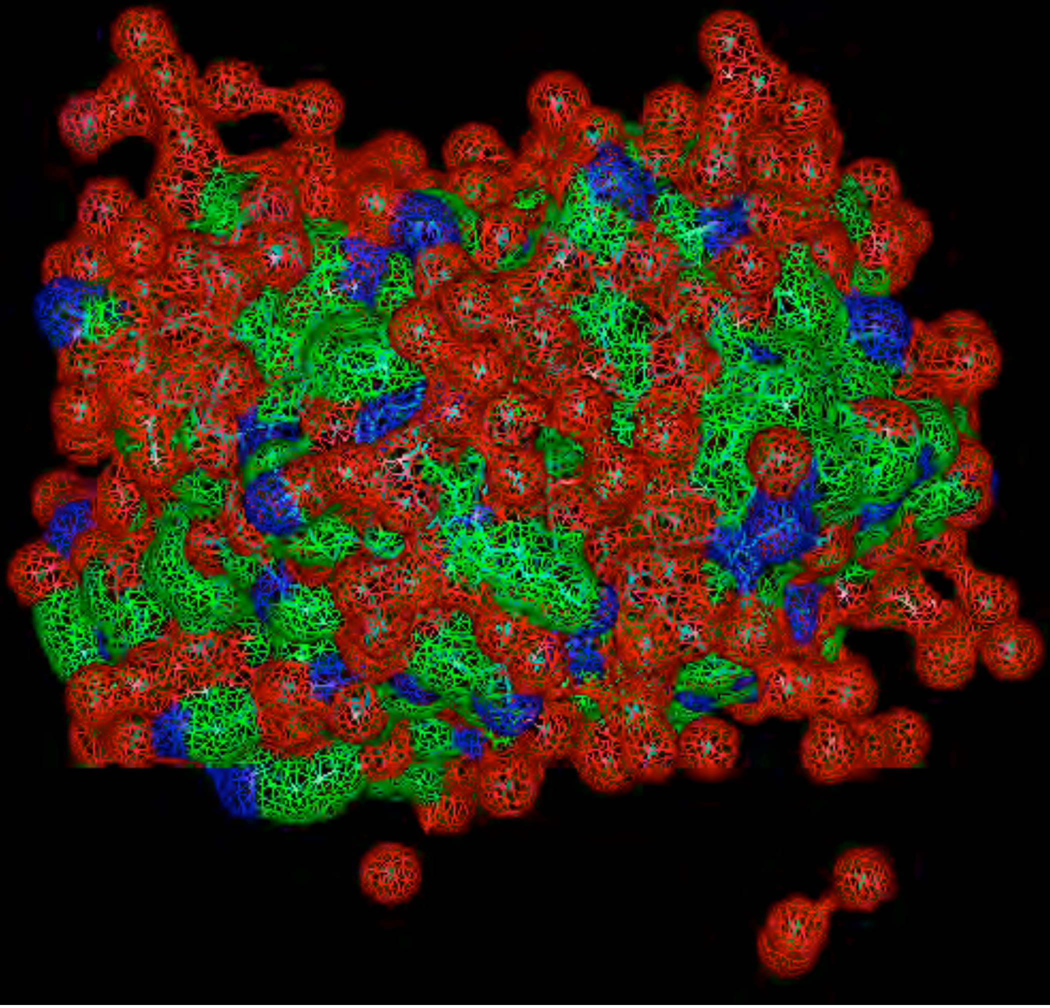
```
HEADER    DNA-BINDING PROTEIN                     20-MAY-94   1ENH      1ENH   2
COMPND    ENGRAILED HOMEODOMAIN                                         1ENH   3
SOURCE    (DROSOPHILA MELANOGASTER) RECOMBINANT FORM EXPRESSED IN       1ENH   4
SOURCE   2 (ESCHERICHIA COLI)                                           1ENH   5
AUTHOR    N.D.CLARKE,C.R.KISSINGER,J.DESJARLAIS,G.L.GILLILAND,C.O.PABO  1ENH   6
REVDAT   1   31-AUG-94 1ENH    0                                        1ENH   7
SCALE3      0.000000  0.000000  0.008466        0.00000                 1ENH  61
ATOM      1  N   ARG     3       2.937  44.573  53.291  1.00 62.68      1ENH  62
ATOM      2  CA  ARG     3       3.220  44.968  51.871  1.00 61.88      1ENH  63
ATOM      3  C   ARG     3       1.922  45.475  51.229  1.00 62.67      1ENH  64
ATOM      4  O   ARG     3       0.984  44.702  51.017  1.00 65.49      1ENH  65
ATOM      5  CB  ARG     3       3.758  43.763  51.101  1.00 58.73      1ENH  66
ATOM      6  CG  ARG     3       3.642  43.884  49.610  1.00 57.06      1ENH  67
ATOM      7  CD  ARG     3       3.776  42.528  48.965  1.00 54.58      1ENH  68
ATOM      8  NE  ARG     3       5.083  42.365  48.340  1.00 56.07      1ENH  69
ATOM      9  CZ  ARG     3       6.183  41.961  48.980  1.00 57.06      1ENH  70
ATOM     10  NH1 ARG     3       6.141  41.670  50.274  1.00 57.63      1ENH  71
ATOM     11  NH2 ARG     3       7.335  41.841  48.325  1.00 57.77      1ENH  72
```

Fichier PDB

- Pas d'information explicite sur les liaisons entre atomes : c'est déterminé par le calcul des distances interatomiques

- Pas d'information sur la charge des atomes: peut être calculé

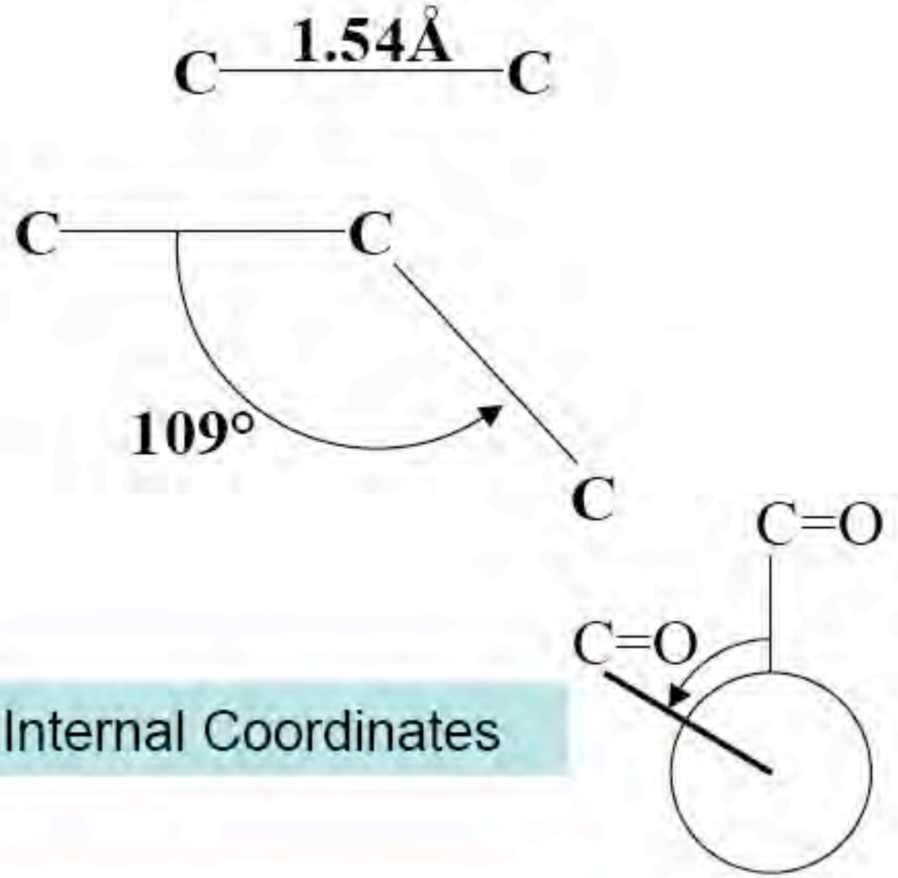
Proteins carry water around with them.



Coordonnées internes

- Indépendantes du référentiel

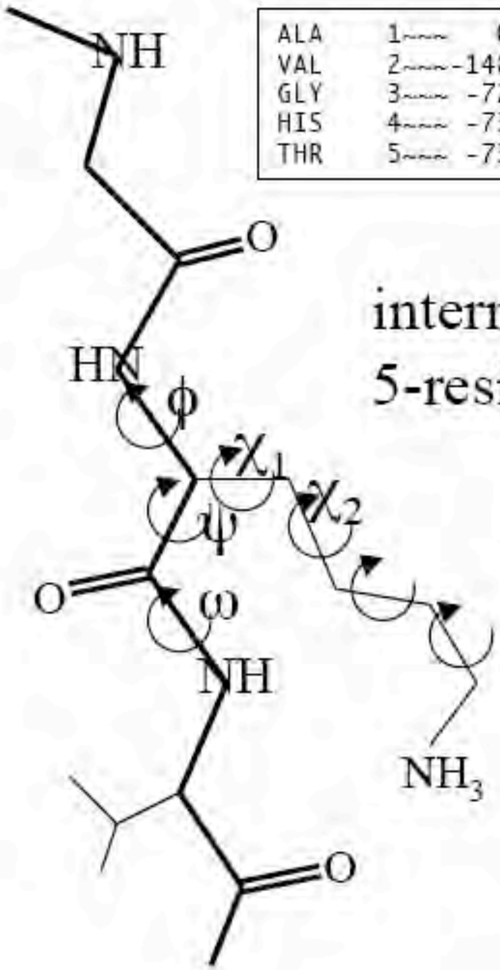
- Components:
  - bond lengths
  - bond angles
  - torsion (dihedral) angles
  - planar groups
  - pairwise distances



NMR structures are solved in Internal Coordinates

Coordonnées internes

Short peptides can be expressed as a set of torsion angles

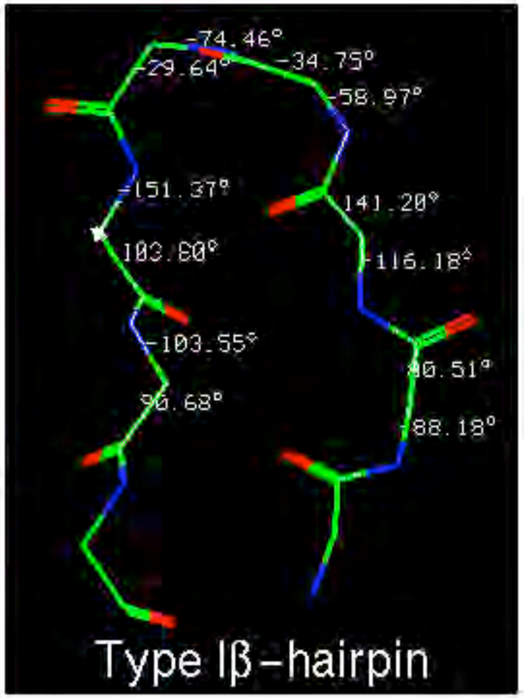


$\phi \quad \psi \quad \omega \quad \chi_1 \quad \chi_2$

```
ALA     1~~~    0.000 127.140 180.000
VAL     2~~~-148.378 111.409 180.000-179.551
GLY     3~~~ -72.763  39.684 180.000
HIS     4~~~ -73.084 122.882 180.000  -87.256 -62.962
THR     5~~~ -73.735 116.210 180.000   49.292
```

internal coordinates of a 5-residue peptide

Cartesian coordinates



Type Iβ-hairpin

If there are sufficient constraints, then internals coordinates may be converted to Cartesian coordinates.

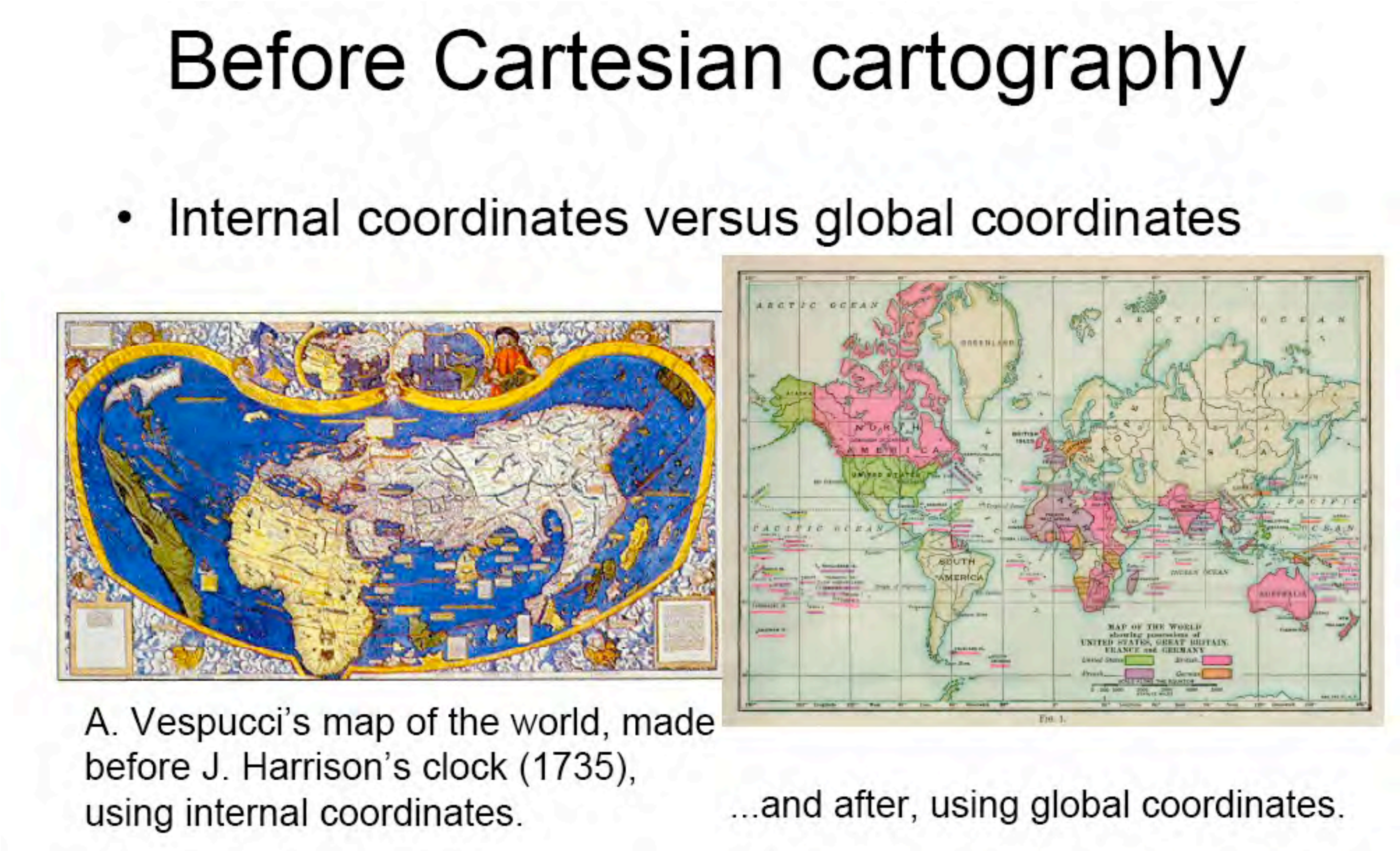
Coordonnées internes et absolues

- Coordonnées internes: au pont, tourner à droite, puis suivre tout droit pendant 500 mètres, tourner à gauche, et continuer pendant 600 mètres

- Coordonnées absolues :

N42° 37′ 04″
W73° 44′ 24″

Coordonnées internes et absolues



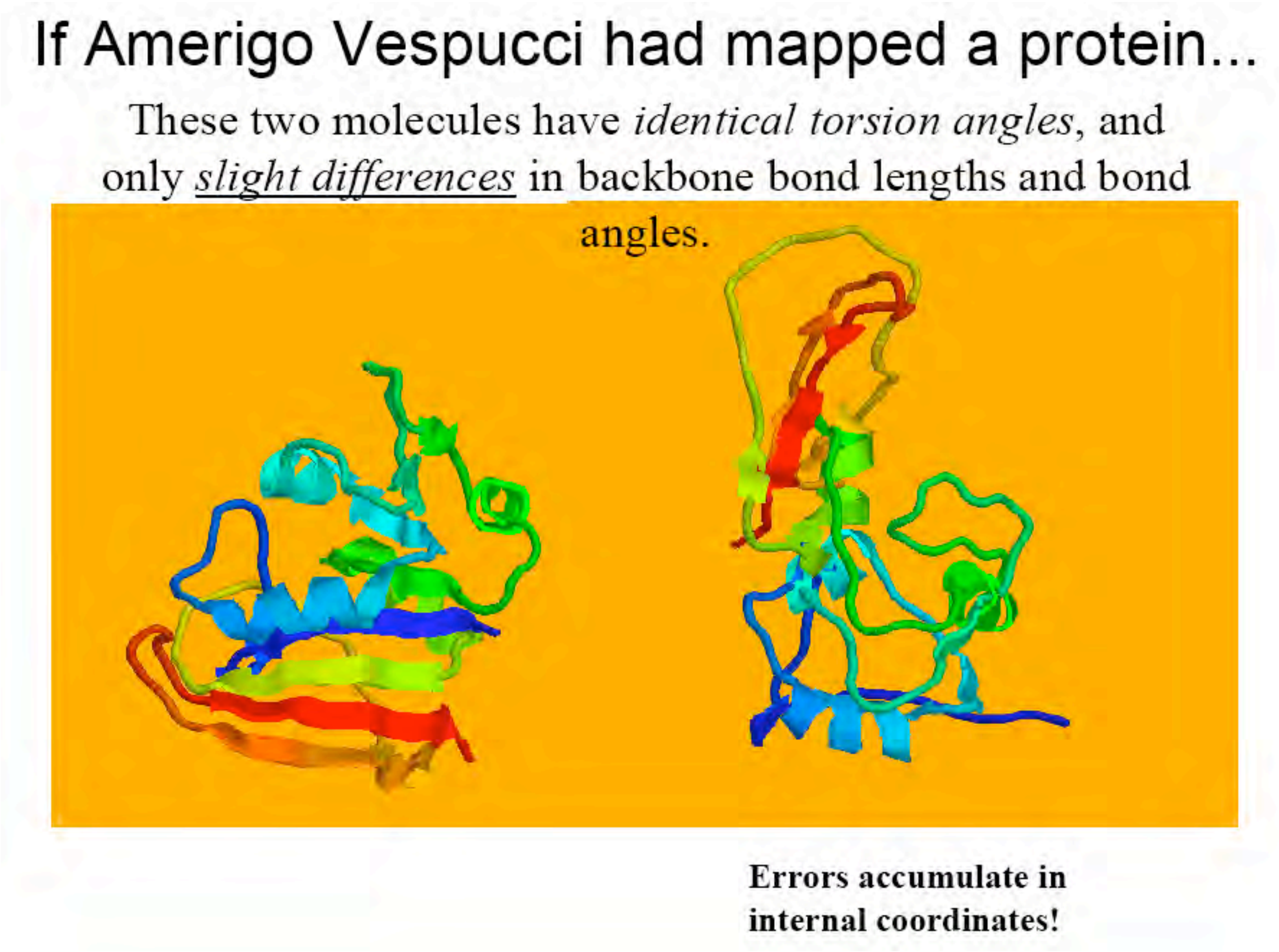
Before Cartesian cartography

- Internal coordinates versus global coordinates

A. Vespucci's map of the world, made before J. Harrison's clock (1735), using internal coordinates.

...and after, using global coordinates.

Coordonnées internes et absolues



If Amerigo Vespucci had mapped a protein...

These two molecules have *identical torsion angles*, and only *slight differences* in backbone bond lengths and bond angles.

Errors accumulate in internal coordinates!

Classification structurale

- **CATH** Class, Architecture, Topology, Homology
  - http://www.cathdb.info/latest/index.html

- **SCOP** Structural Classification of Proteins
  - http://scop.mrc-lmb.cam.ac.uk/scop/

- **DALI** Structural Classification of Proteins
  - http://www.ebi.ac.uk/dali/

- **PDBSUM** Attribution des structures secondaires à partir du code PDB
  - http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/

Taxinomie structurale CATH

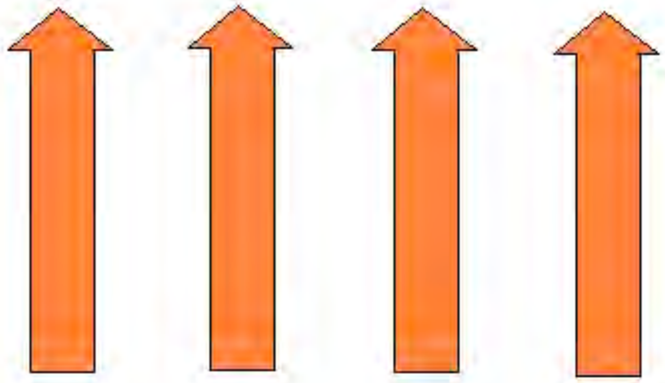
CLASSES

- **alpha** (**principalement** hélices); pas tout alpha
- **beta** (**principalement** brins)
- **alpha+beta**, dissociation des domaines α et β
- **alpha/beta**, alternance, unités βαβ
- **few**, séquence courte, stabilisation non effectuée par les SSR mais par les ponts disulfure

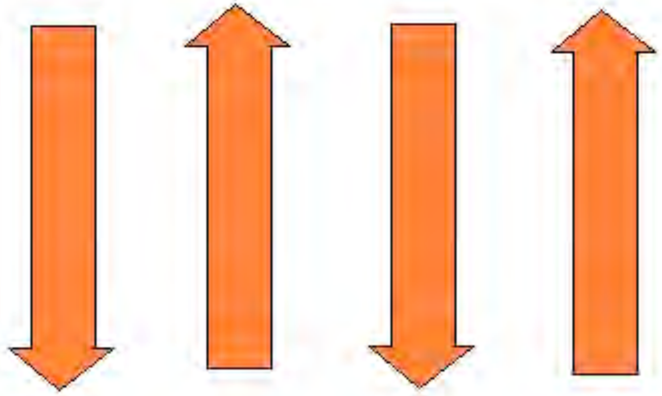
Consensus de 3 méthodes automatiques de délimitations de domaines, le reste manuellement
Homologie définie à partir de l'identité de séquences

Examples of protein architecture

Architecture refers to the arrangement and orientation of SSEs, but not to the connectivity.



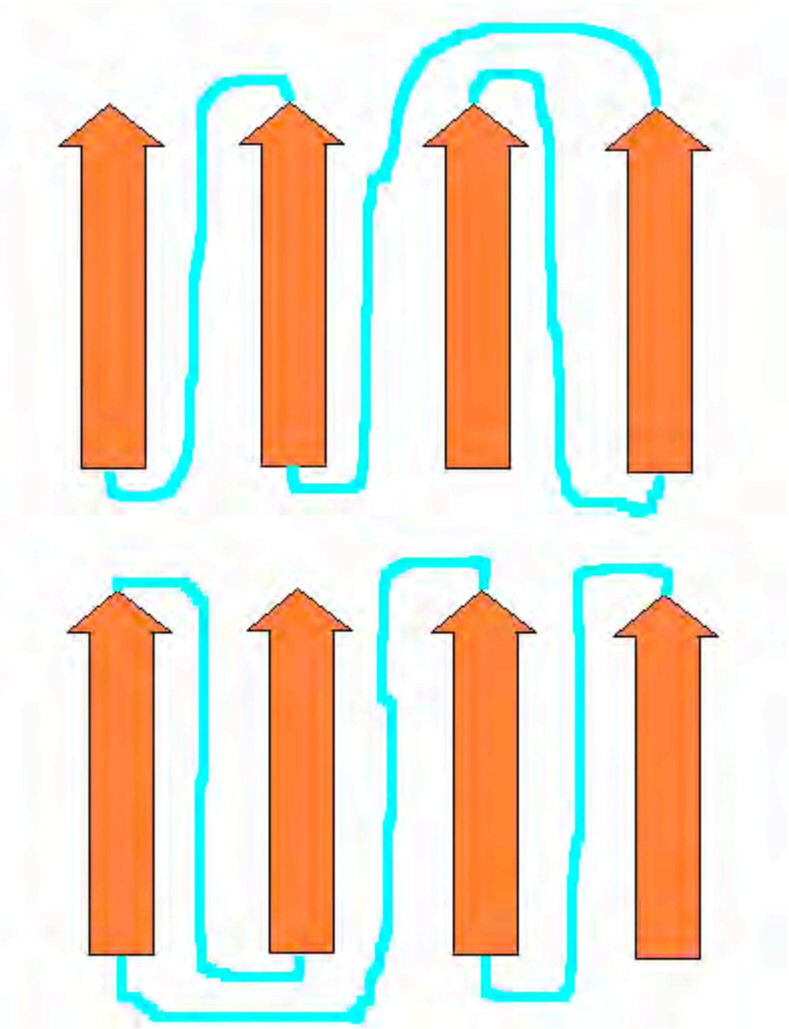
β-sheet with all pairs of strands parallel



β-sheet with all pairs of strands anti-parallel

Examples of protein topology

Topology refers to the manner in which the SSEs are connected.



Two β-sheets (all parallel) with different topologies.

CATH

http://www.cathdb.info/

● PDB Code
○ CATH Code
○ General Text

Home > Top > Class1 > 10 > 210 > 10 > 1 > 1 > 1 > 1utg00 View this page as XML

CATH Domain 1utg00

Classification

Class Mainly Alpha	1
Architecture Orthogonal Bundle	1.10
Topology Uteroglobin	1.10.210
Homologous Superfamily Uteroglobin	1.10.210.10
Sequence Family (S35) Uteroglobin	1.10.210.10.1
Non-identical (S95) Uteroglobin	1.10.210.10.1.1
Identical (S100) Uteroglobin	1.10.210.10.1.1.1

PDB Information

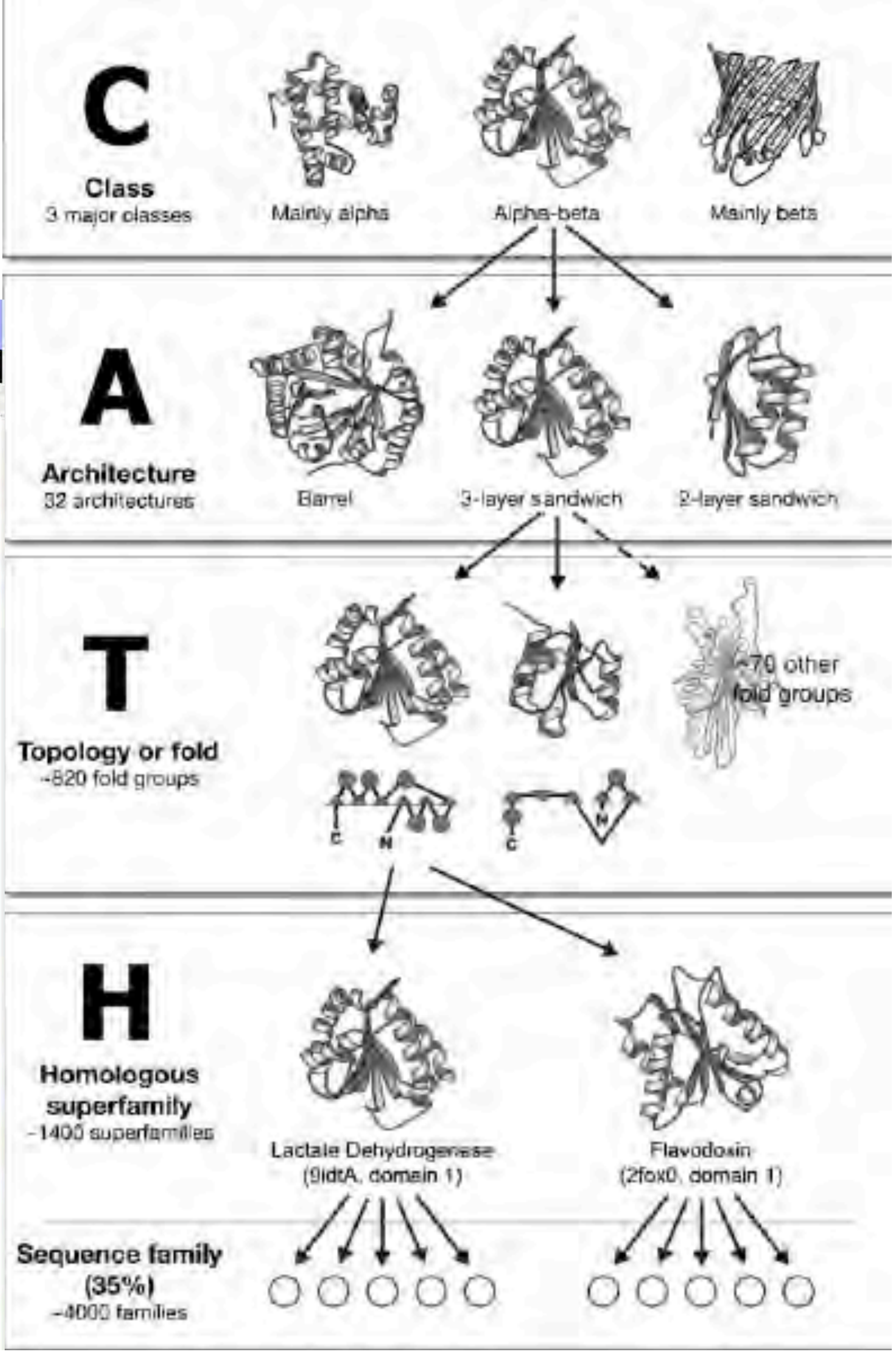
PDB Code	1utg
PDB Header	Uteroglobin (oxidized)
PDB Source	Rabbit (oryctolagus cuniculus) female genital tract

Domain Information

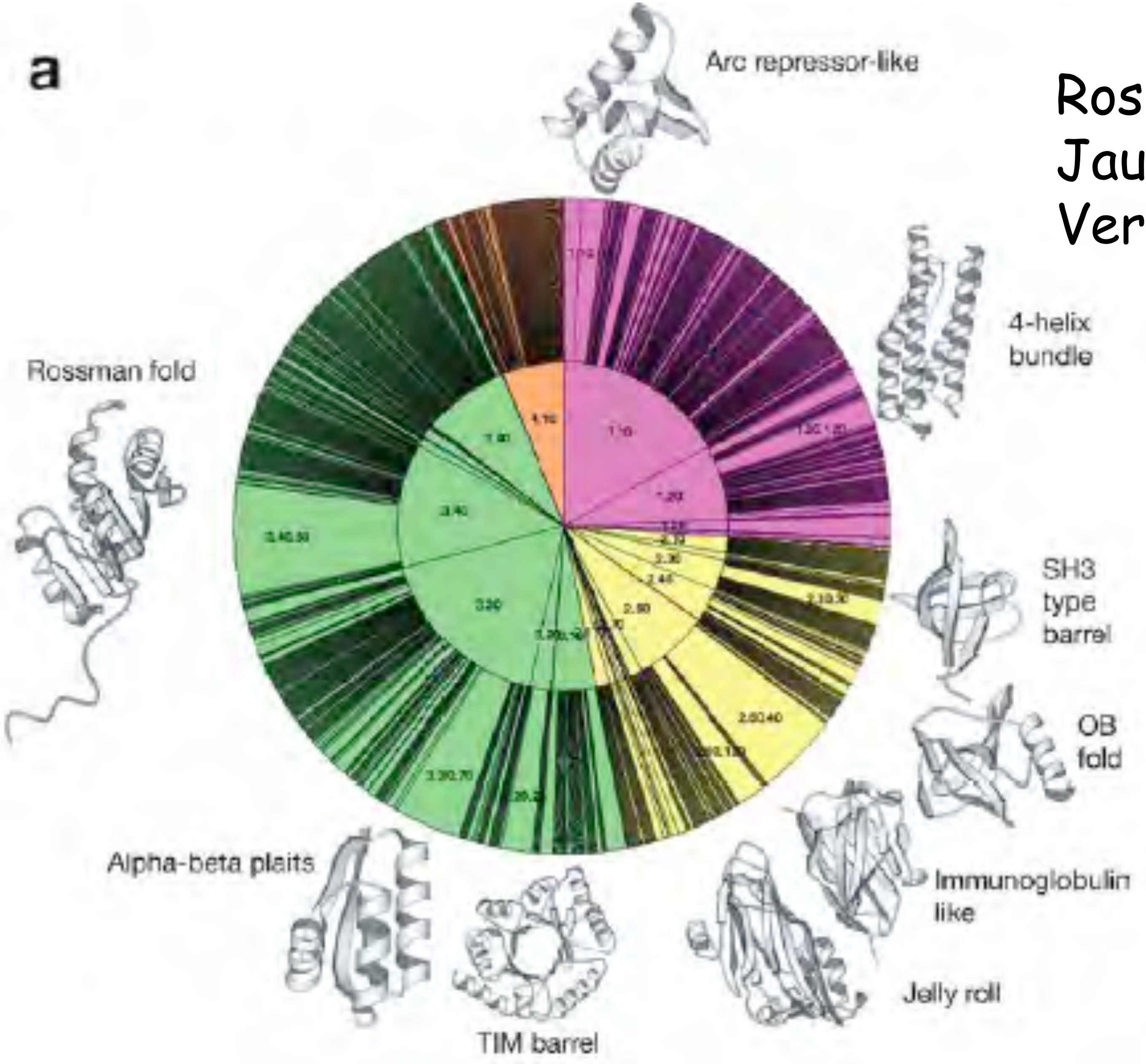
Domain Sequence	GICPRFAHVIENLLLGTPSSYETSLKEFEPDDTMKDAGM QMKKVLDSLPQTTRENIMKLTEKIVKSPLCM

CATH considers structural domains as semi-independent folding units. It is quite common that a structural domain is made up of more than one sequence segment (*i.e.* non-sequential stretch of peptide). The table below provides information for each of the segments that make up the structural domain.

Segment Number	Segment Length	PDB Start	PDB Stop
1	70	1	70



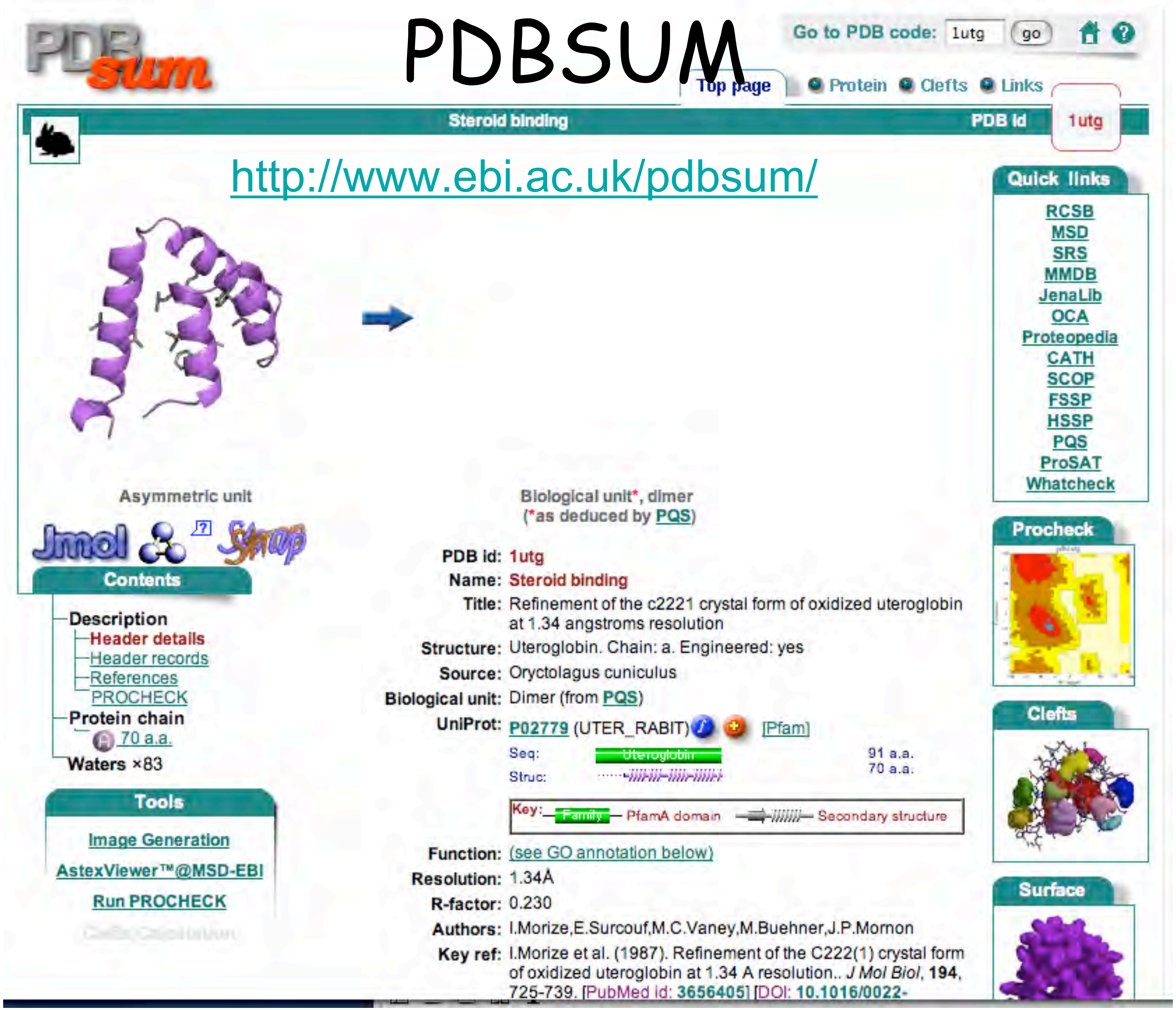
Roue CATHerine



Rose :Alpha
Jaune : Beta
Vert : Alpha-beta

PDBSUM

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



PDB id: 1utg

Name: Steroid binding

Title: Refinement of the c2221 crystal form of oxidized uteroglobin at 1.34 angstroms resolution

Structure: Uteroglobin. Chain: a. Engineered: yes

Source: Oryctolagus cuniculus

Biological unit: Dimer (from PQS)

UniProt: P02779 (UTER_RABIT) [Pfam]

Seq: Uteroglobin 91 a.a.

Struc: 70 a.a.

Key: Family PfamA domain Secondary structure

Function: (see GO annotation below)

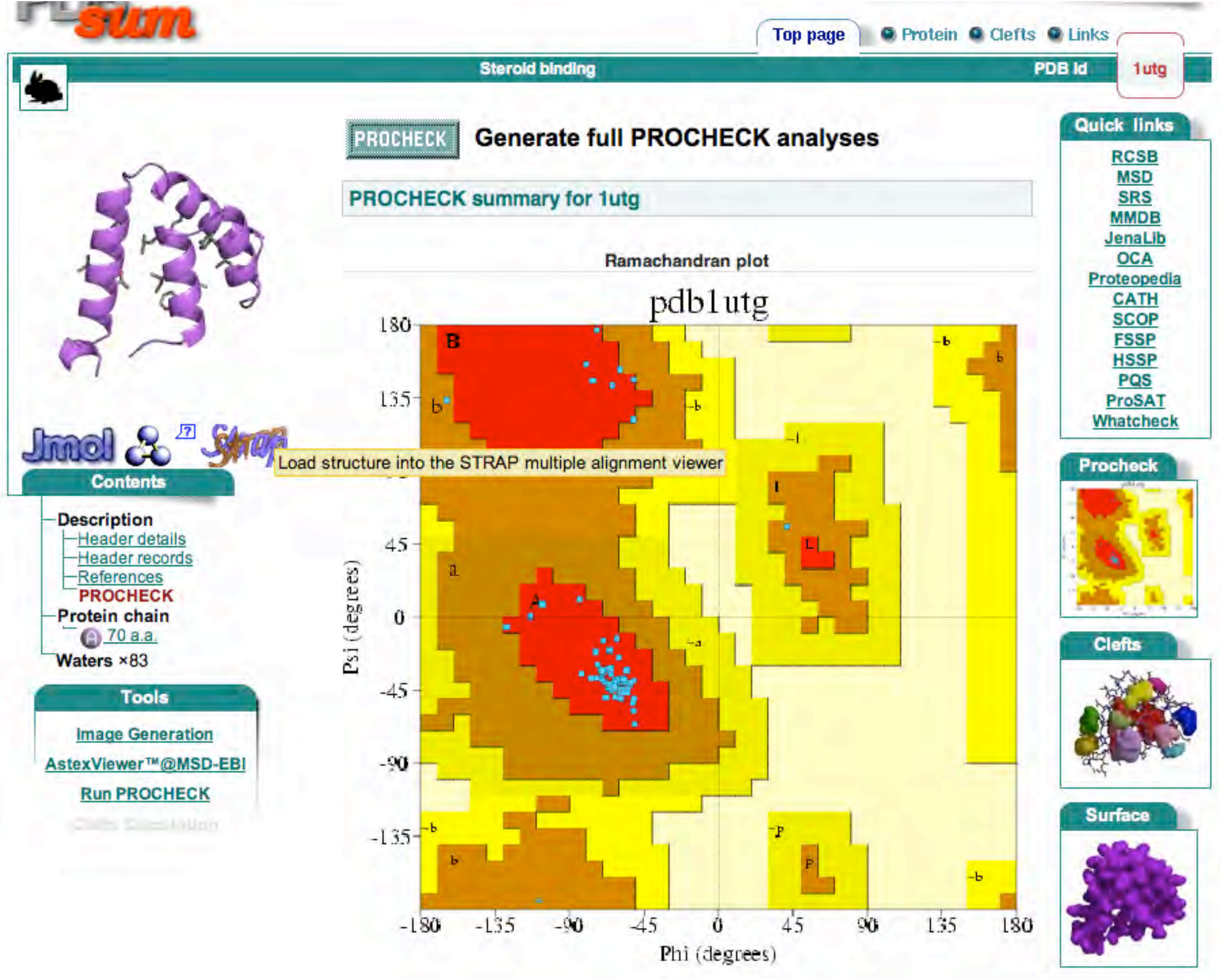
Resolution: 1.34Å

R-factor: 0.230

Authors: I.Morize,E.Surcouf,M.C.Vaney,M.Buehner,J.P.Mornon

Key ref: I.Morize et al. (1987). Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 A resolution.. *J Mol Biol*, **194**, 725-739. [PubMed id: 3656405] [DOI: 10.1016/0022-

PDBSUM



Generate full PROCHECK analyses

PROCHECK summary for 1utg

Load structure into the STRAP multiple alignment viewer

SCOP

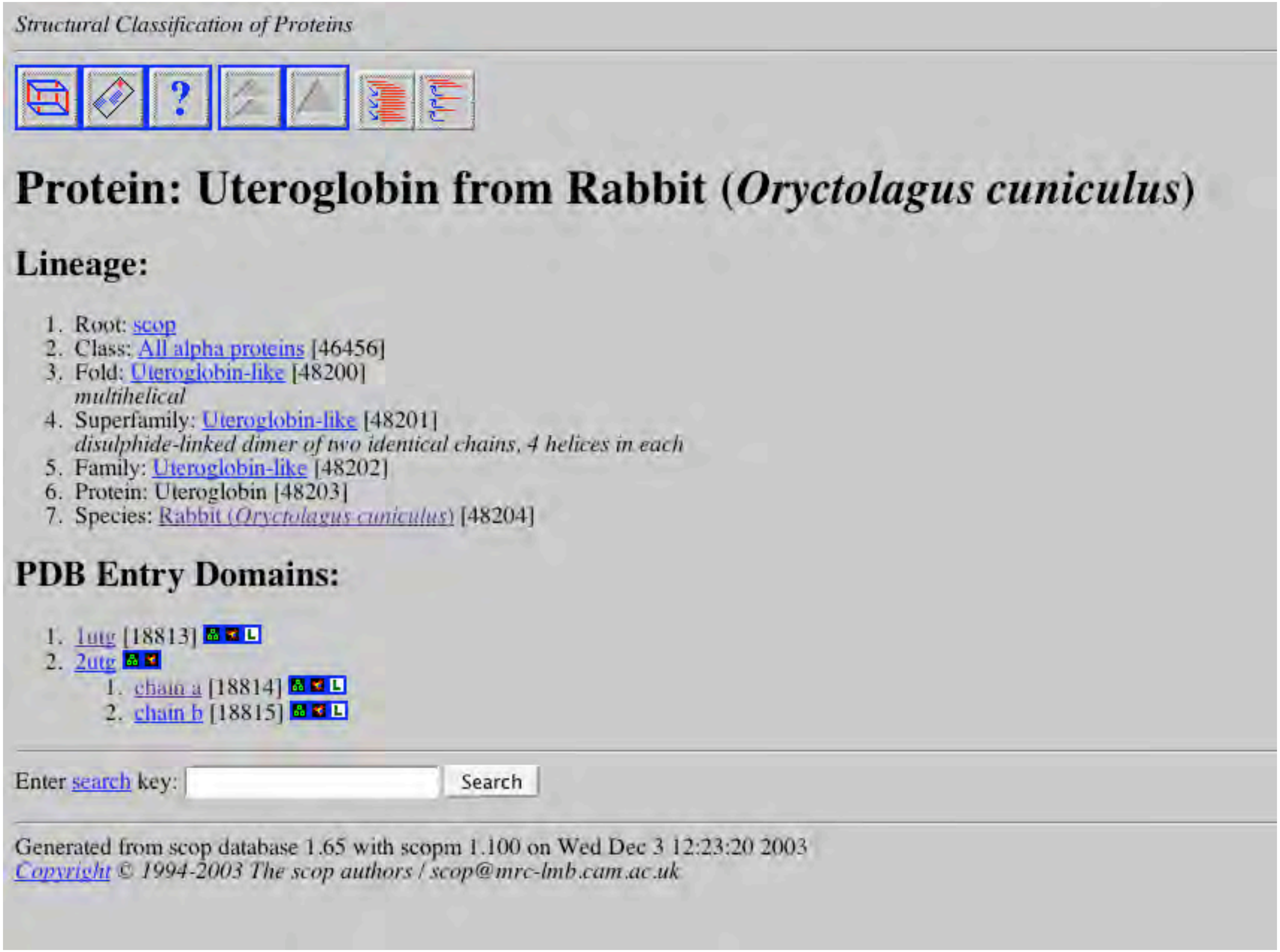
Classes : comme CATH

Repliements : même repliement si même arrangement des SSR et même connexion topologique (cœur commun + éléments de structures secondaires périphériques + régions de turns qui diffèrent en taille et en conformation)

Superfamilles : identités de séquence faibles; détails structuraux et fonction suggèrent ancêtre commun

Familles : relations évolutives par similitudes de séquence significatives, ou similitudes structure et fonction

Essentiellement manuel



Structural Classification of Proteins

Protein: Uteroglobin from Rabbit (*Oryctolagus cuniculus*)

Lineage:

1. Root: scop
2. Class: All alpha proteins [46456]
3. Fold: Uteroglobin-like [48200]
   multihelical
4. Superfamily: Uteroglobin-like [48201]
   disulphide-linked dimer of two identical chains, 4 helices in each
5. Family: Uteroglobin-like [48202]
6. Protein: Uteroglobin [48203]
7. Species: Rabbit (*Oryctolagus cuniculus*) [48204]

PDB Entry Domains:

1. 1utg [18813] L
2. 2utg
    1. chain a [18814] L
    2. chain b [18815] L

Enter search key: Search

Generated from scop database 1.65 with scopm 1.100 on Wed Dec 3 12:23:20 2003

SCOP

1 - Classe: caractéristiques globales, pas de relation évolutive

2 - Repliement: topologie similaire. Cousins éloignés d'un point de vue évolutif?

3 - Superfamille : homologie structurale claire

4 - Famille: homologie de séquence claire

5 - Protéine : identité fonctionnelle

6 - Espèces : séquences uniques

SCOP

	2002	2008
Repliements tout alpha	151	259
Repliements tout beta	110	165
Repliements alpha/beta	113	141
Repliements alpha+beta	208	334
Multi domaines	34	53
Protéines membranaires	12	50
Petites protéines	58	85
Total de repliements	686	1086

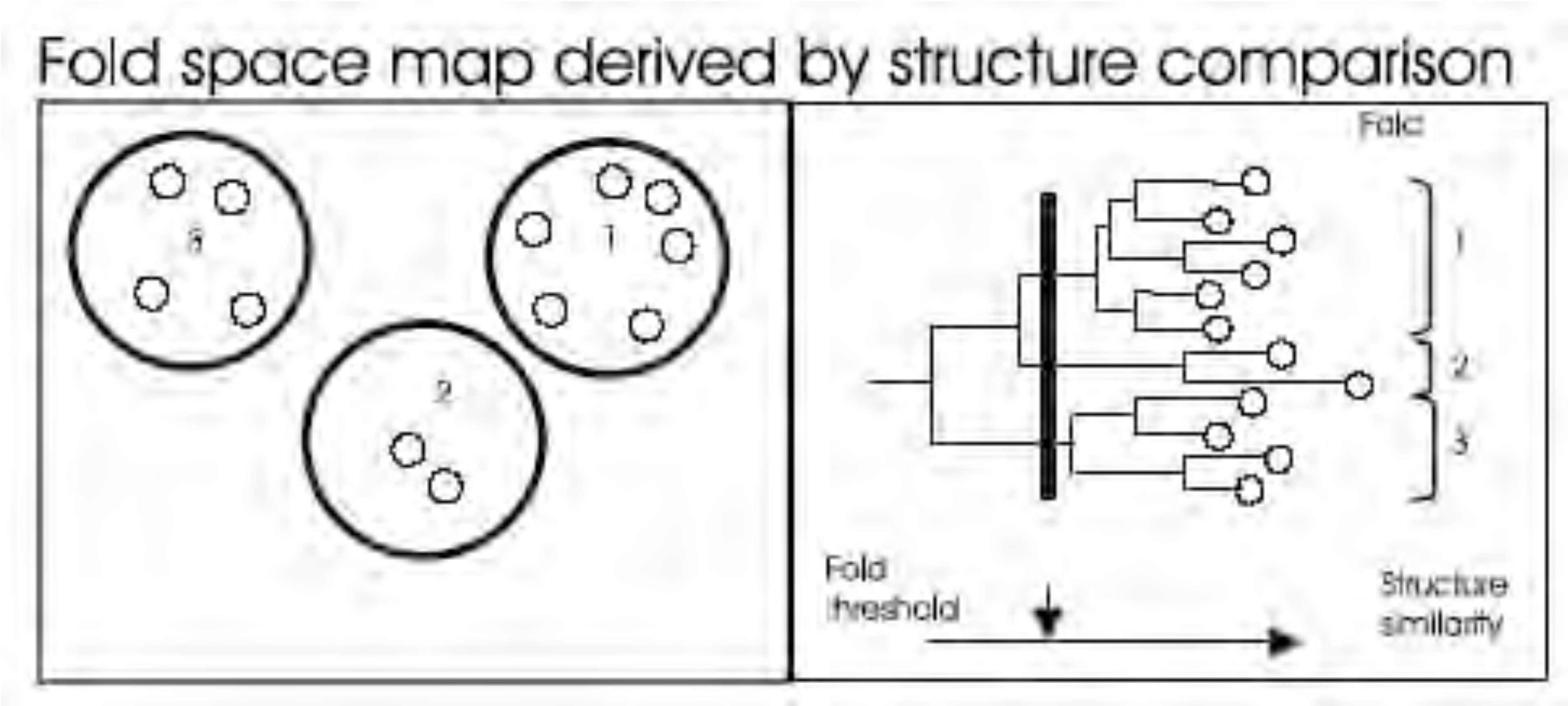
DALI

Méthode automatique (réseau de neurones et tri hiérarchique) : identifie dans une première étape les similitudes structurales: carte de l'espace des repliements (ressemblance structurale suffisante pour opérer les différences entre repliements)

Autre critère pour définir les superfamilles : déduit des propriétés fonctionnelles

Dietman (2001) NSB 8:953

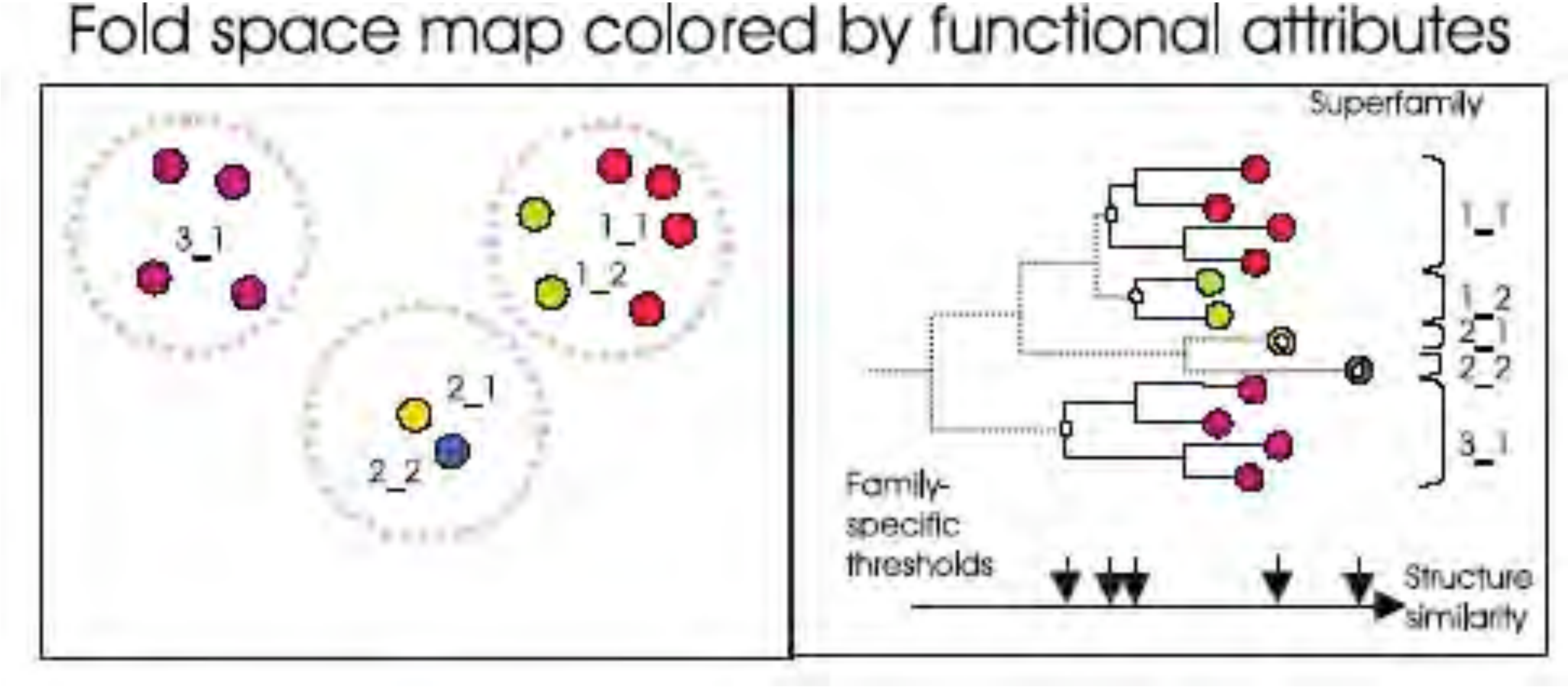
DALI



Cercles épais = repliements ; déduits des similitudes structurales

Dietman (2001) NSB 8:953

DALI

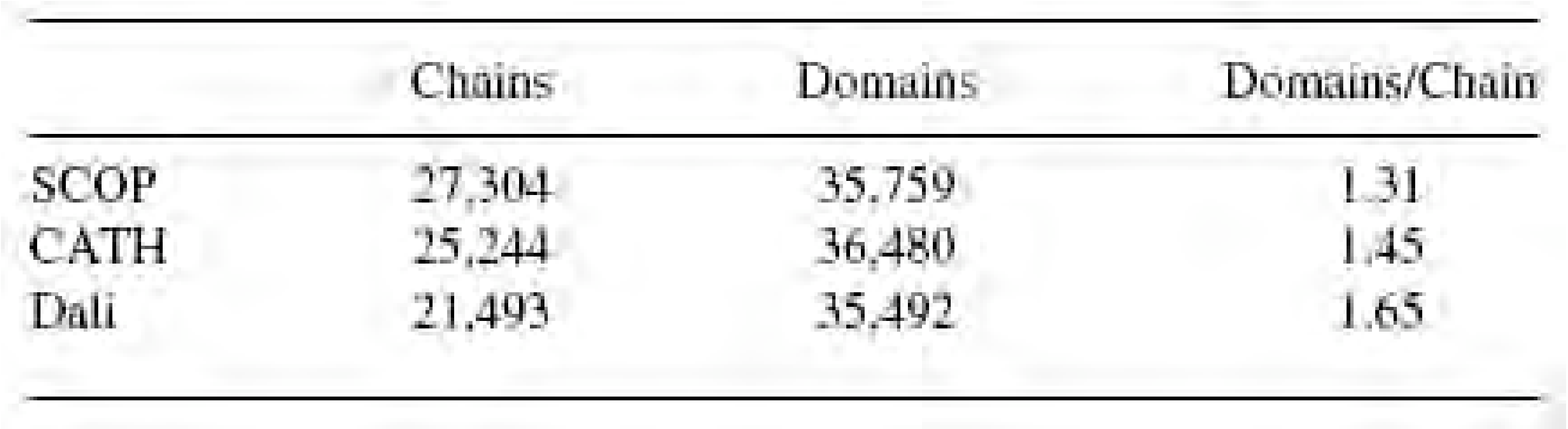


Attribution des superfamilles à partir des fonctions

Dietman (2001) NSB 8:953

Consensus des classes

SCOP considéré comme standard, en absence d'homologie avec un autre domaine, classification manuelle
CATH = consensus de 3 méthodes automatiques de délimitations de domaines, le reste manuellement. Homologie définie à partir de l'identité de séquences.
Dali = définition des domaines entièrement automatique (repose sur l'algorithme PUU)



	Chains	Domains	Domains/Chain
SCOP	27,304	35,759	1.31
CATH	25,244	36,480	1.45
Dali	21,493	35,492	1.65

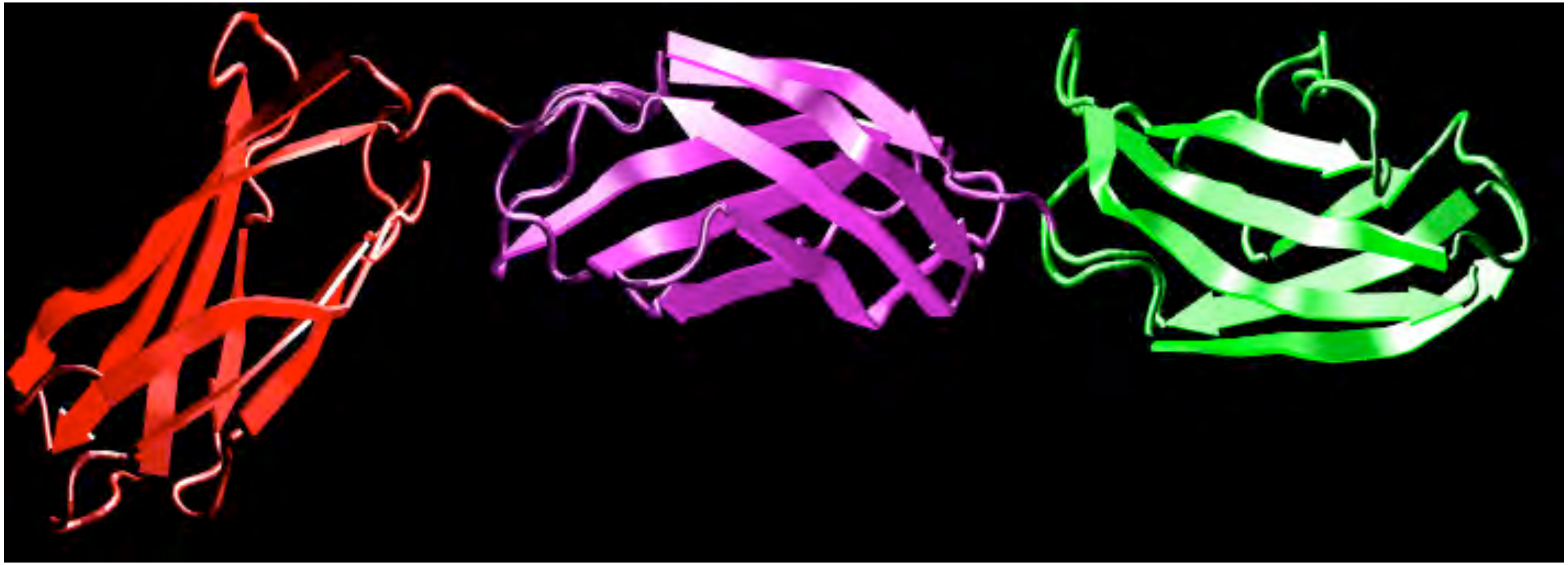
Statistiques du nombre de chaînes, de domaines tels que définis par SCOP, CATH et Dali
Day Prot Sci (2003) 12:2150

Domaines

Domaines globulaires

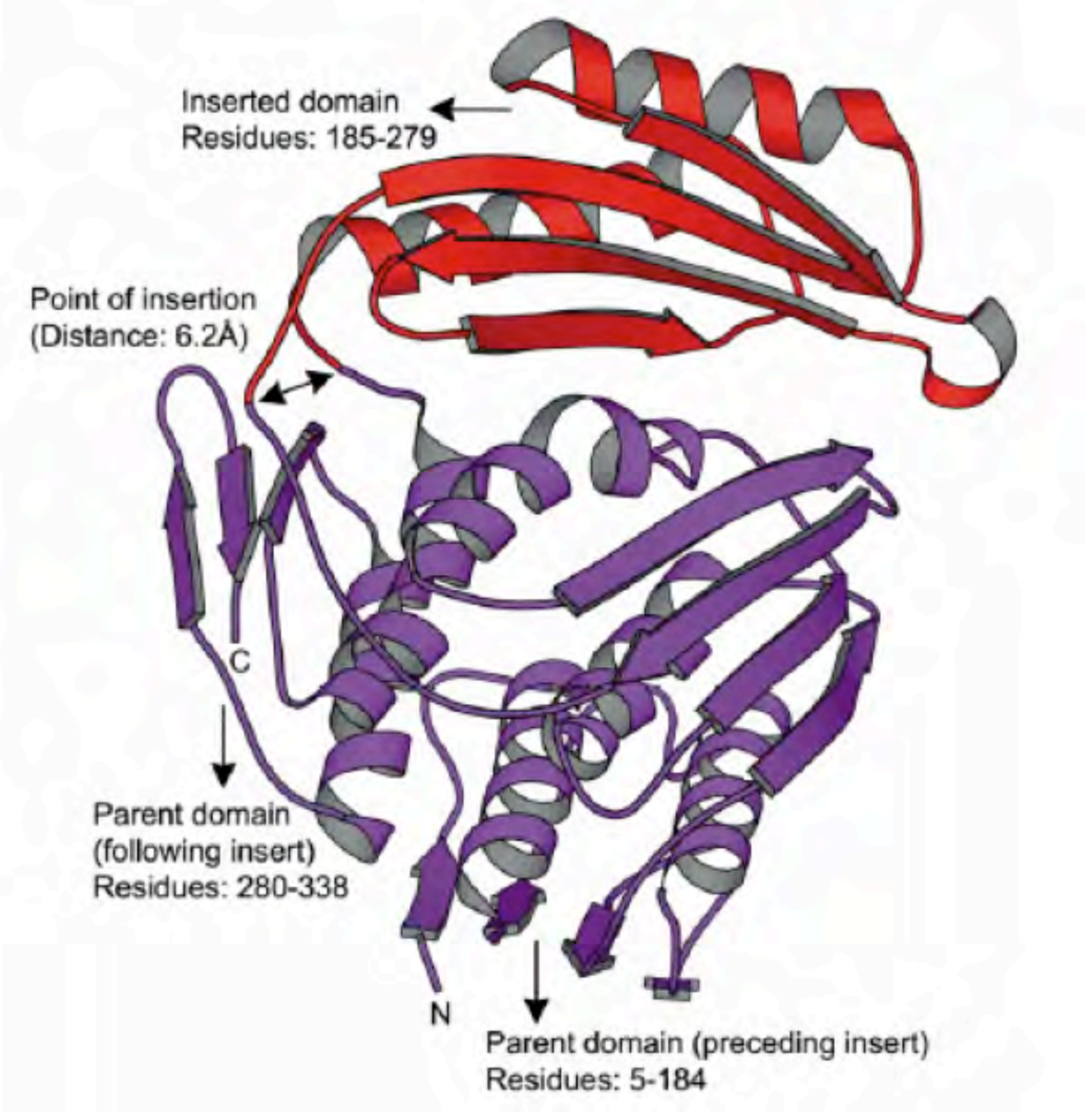
- Une chaîne polypeptidique peut comprendre plusieurs domaines, parfois appelés lobes, comme dans la Béta galactosidase.
- Un domaine est défini par ses caractéristiques 3D de compacité.
- Il ne correspond pas nécessairement à une continuité dans la séquence.
- Il peut y avoir une fonction par domaine.

Fibronectine

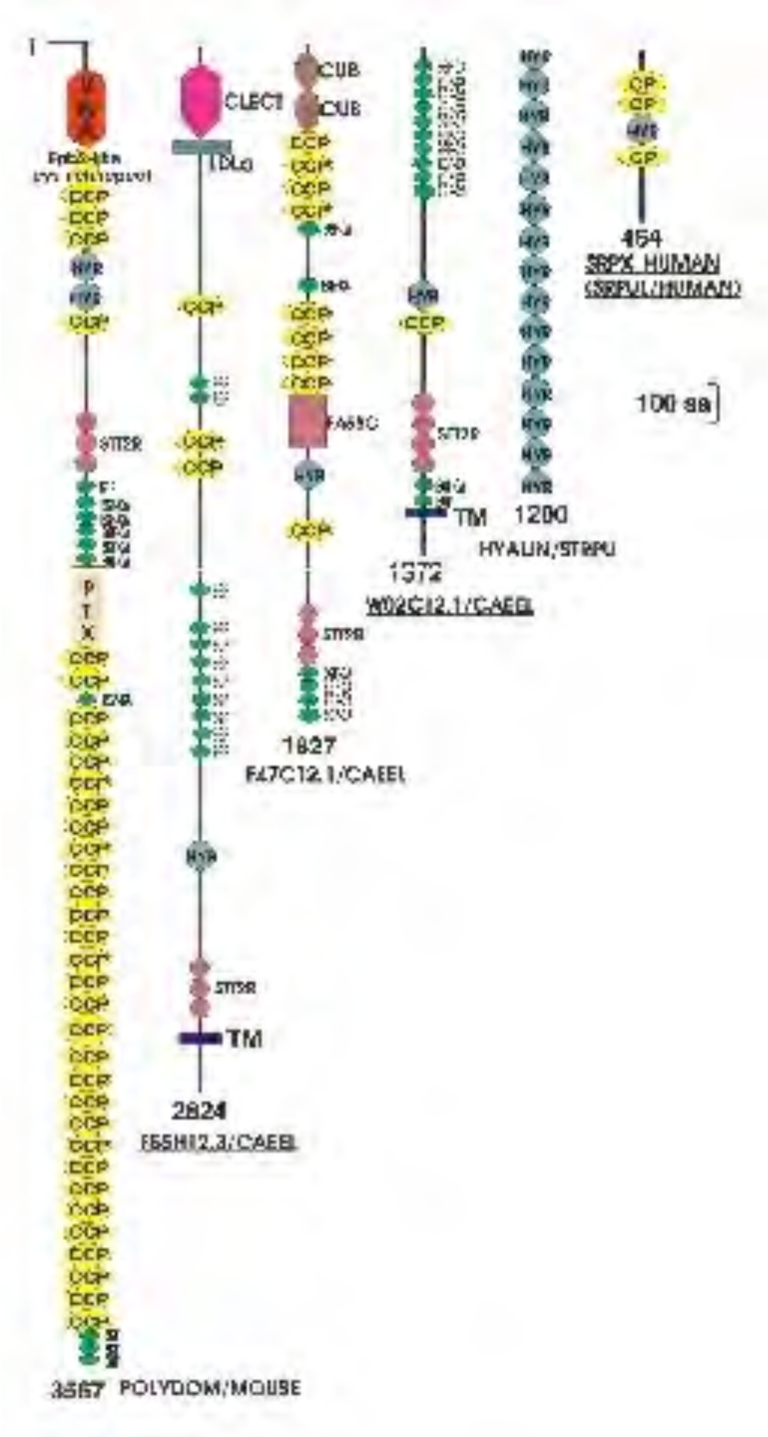


Liaison aux surfaces cellulaires et à la fibronectine,
l 'héparine, l 'actine...

80% domaines = segments contigu d'une seule chaîne
Parfois insertion d'un domaine dans un autre (parent)



Domaine type thioredoxine inséré (rouge) dans un domaine parent (bleu) : 1qmhA



Polydom

Pentraxin (PTX)

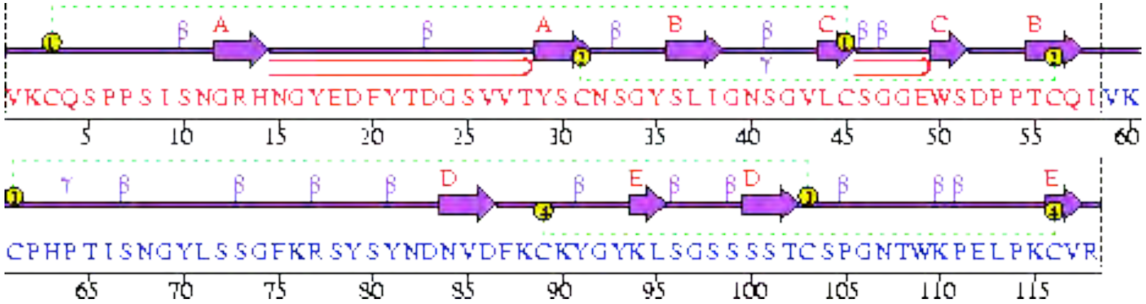
complement control protein (CCP)

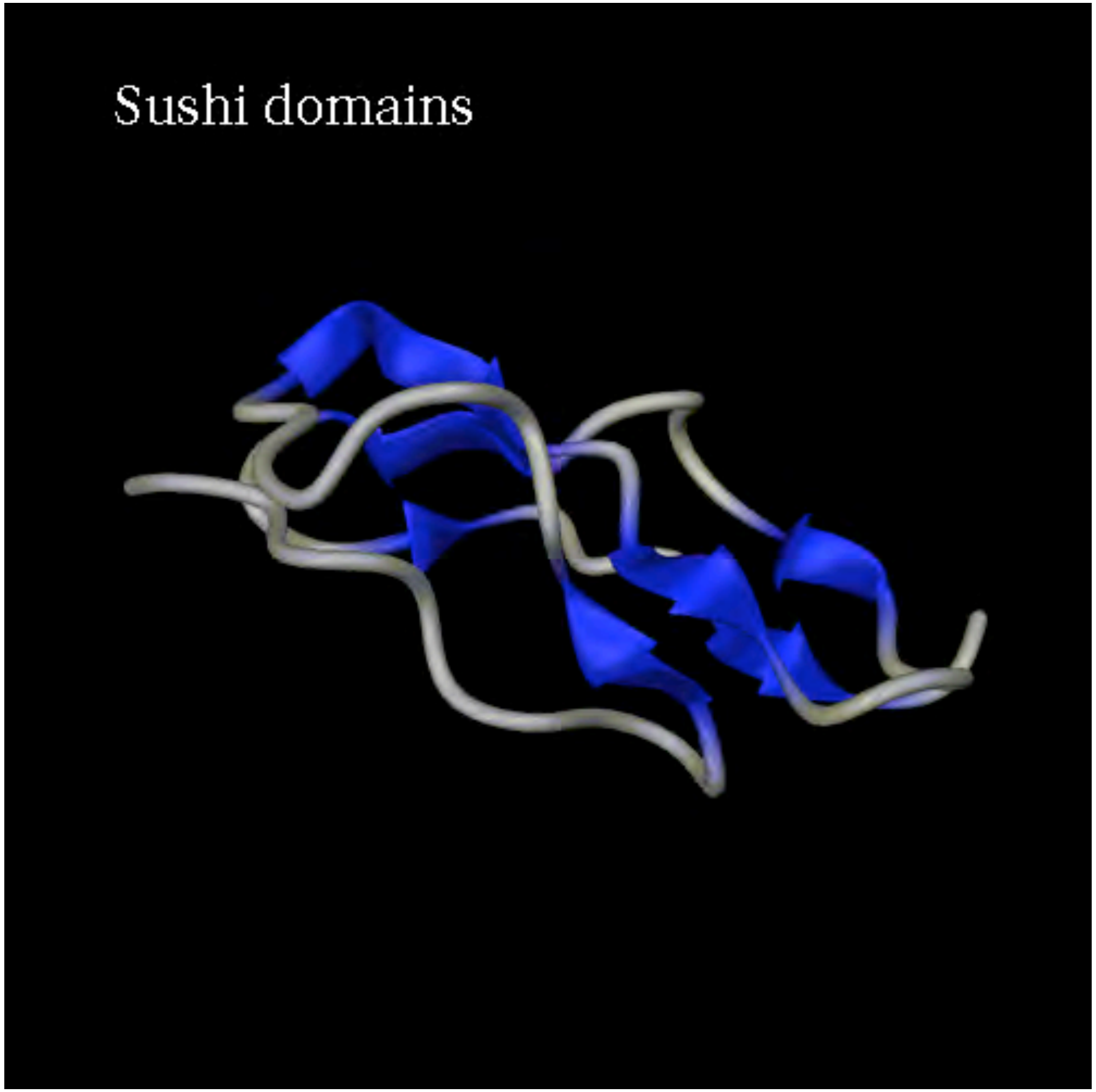
epidermal growth factor (EGF)

von willebrand factor A (VWA)

AF206329

SUSHI





- Intégré à la membrane
- Régulation négative de l'activation du complément
- Protéine de liaison au complément

Banques de domaines

PRODOM
http://prodes.toulouse.inra.fr/prodom/current/html/home.php

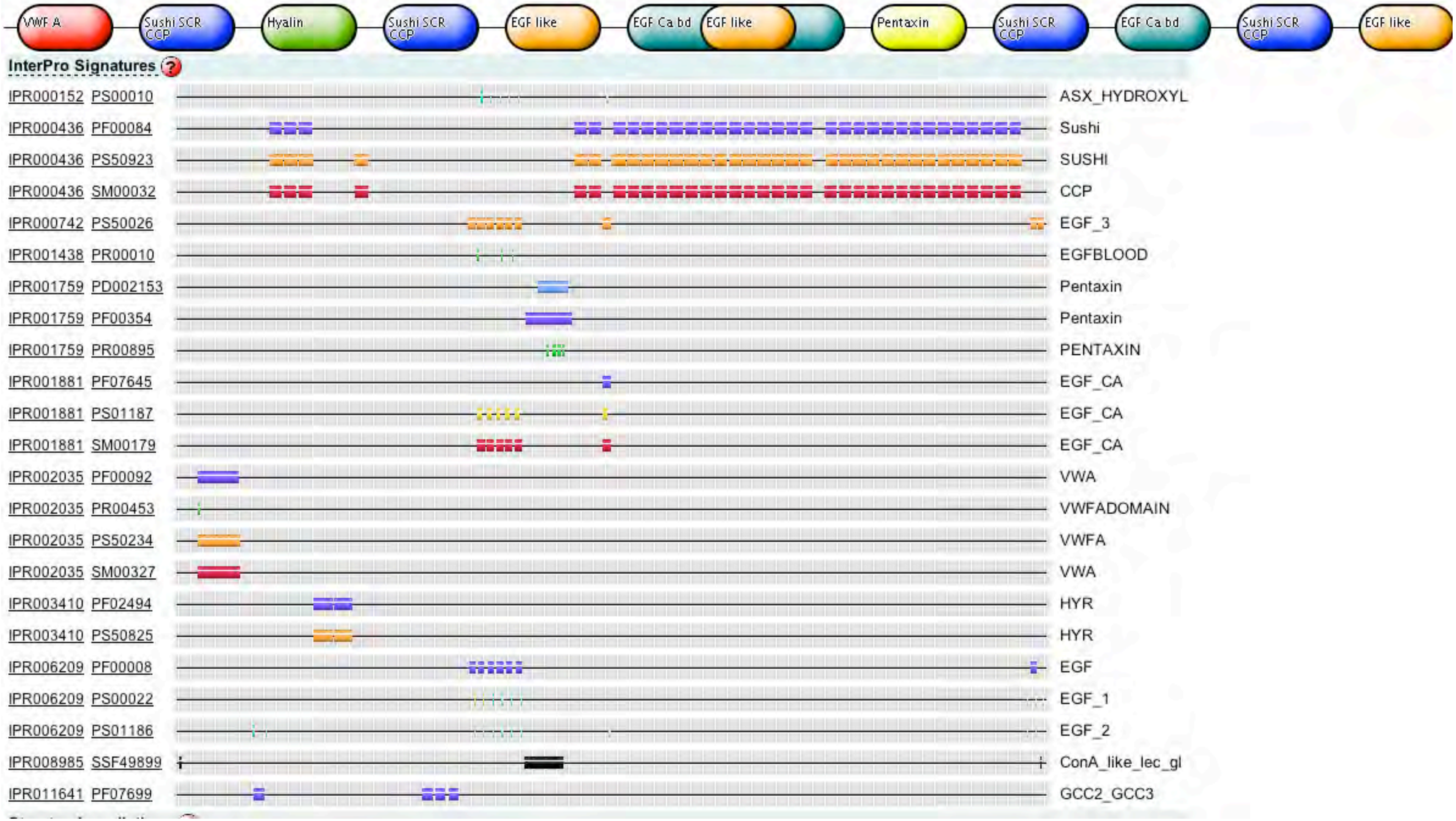
SMART
http://smart.embl-heidelberg.de/

PFAM
http://www.sanger.ac.uk/Software/Pfam/

Interpro
http://www.ebi.ac.uk/interpro/

Domaines Q9ES77_MOUSE

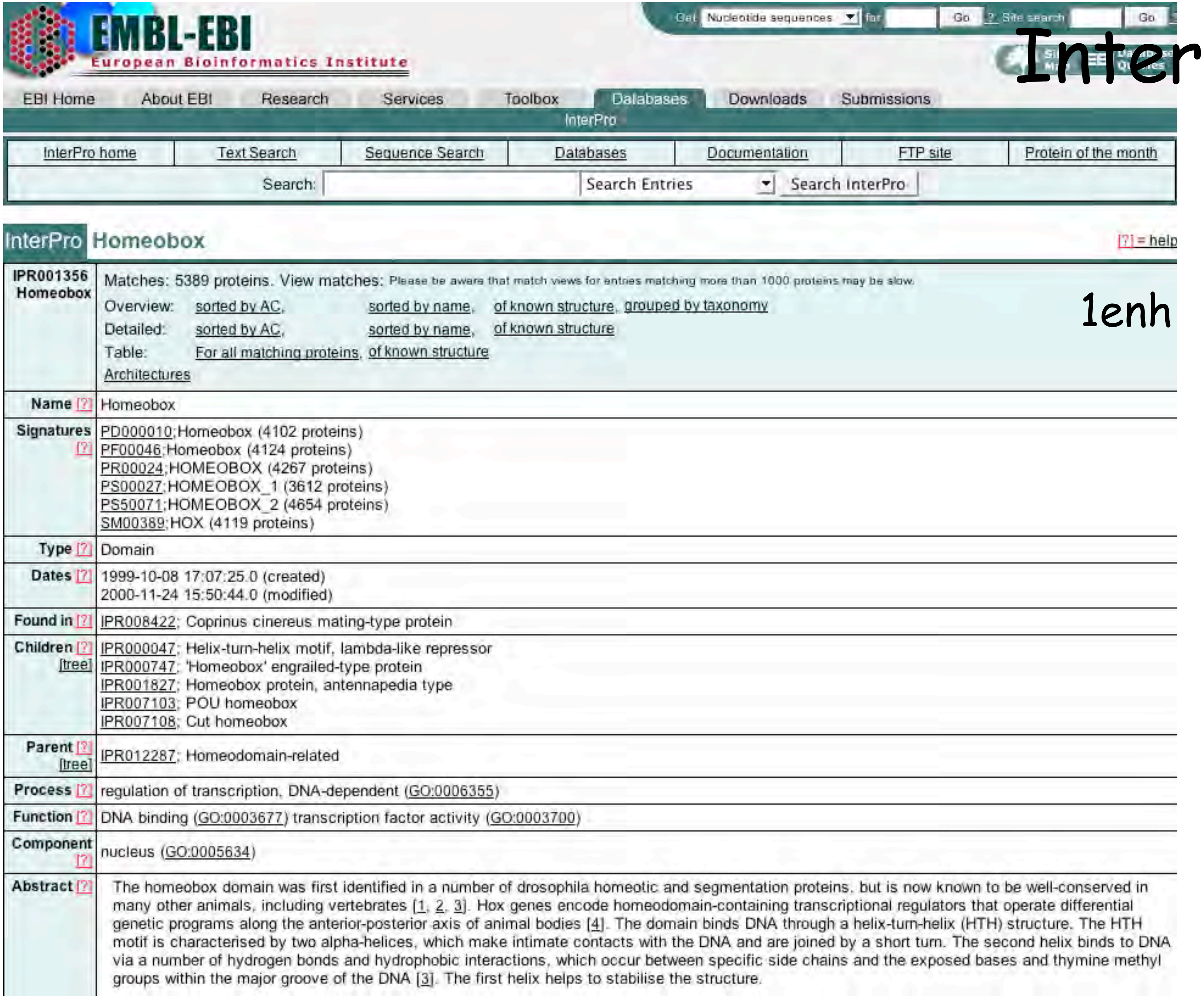
InterPro



InterPro Signatures ?

IPR000152	PS00010	ASX_HYDROXYL
IPR000436	PF00084	Sushi
IPR000436	PS50923	SUSHI
IPR000436	SM00032	CCP
IPR000742	PS50026	EGF_3
IPR001438	PR00010	EGFBLOOD
IPR001759	PD002153	Pentaxin
IPR001759	PF00354	Pentaxin
IPR001759	PR00895	PENTAXIN
IPR001881	PF07645	EGF_CA
IPR001881	PS01187	EGF_CA
IPR001881	SM00179	EGF_CA
IPR002035	PF00092	VWA
IPR002035	PR00453	VWFADOMAIN
IPR002035	PS50234	VWFA
IPR002035	SM00327	VWA
IPR003410	PF02494	HYR
IPR003410	PS50825	HYR
IPR006209	PF00008	EGF
IPR006209	PS00022	EGF_1
IPR006209	PS01186	EGF_2
IPR008985	SSF49899	ConA_like_lec_gl
IPR011641	PF07699	GCC2_GCC3

InterPro

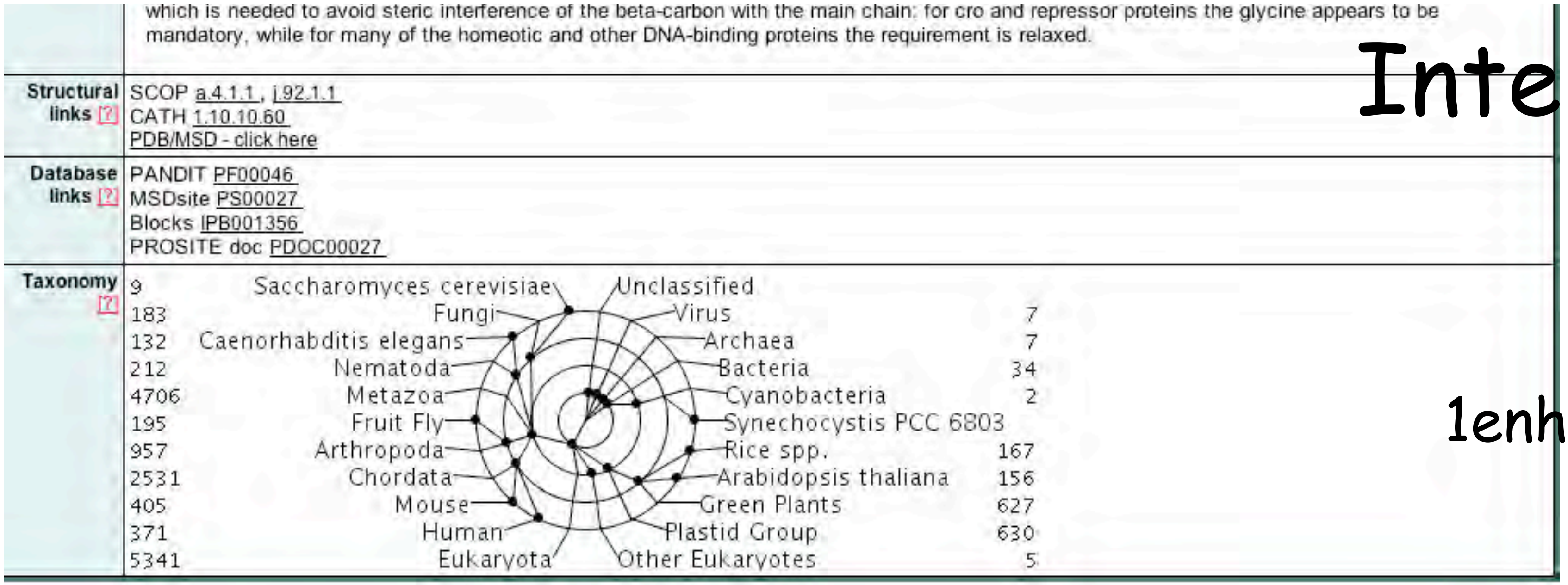


InterPro **Homeobox**

1enh

IPR001356 Homeobox	Matches: 5389 proteins. View matches: Please be aware that match views for entries matching more than 1000 proteins may be slow. Overview: sorted by AC, sorted by name, of known structure, grouped by taxonomy Detailed: sorted by AC, sorted by name, of known structure Table: For all matching proteins, of known structure Architectures
Name [?]	Homeobox
Signatures [?]	PD000010;Homeobox (4102 proteins) PF00046;Homeobox (4124 proteins) PR00024;HOMEOBOX (4267 proteins) PS00027;HOMEOBOX_1 (3612 proteins) PS50071;HOMEOBOX_2 (4654 proteins) SM00389;HOX (4119 proteins)
Type [?]	Domain
Dates [?]	1999-10-08 17:07:25.0 (created) 2000-11-24 15:50:44.0 (modified)
Found in [?]	IPR008422; Coprinus cinereus mating-type protein
Children [?] [tree]	IPR000047; Helix-turn-helix motif, lambda-like repressor IPR000747; 'Homeobox' engrailed-type protein IPR001827; Homeobox protein, antennapedia type IPR007103; POU homeobox IPR007108; Cut homeobox
Parent [?] [tree]	IPR012287; Homeodomain-related
Process [?]	regulation of transcription, DNA-dependent (GO:0006355)
Function [?]	DNA binding (GO:0003677) transcription factor activity (GO:0003700)
Component [?]	nucleus (GO:0005634)
Abstract [?]	The homeobox domain was first identified in a number of drosophila homeotic and segmentation proteins, but is now known to be well-conserved in many other animals, including vertebrates [1, 2, 3]. Hox genes encode homeodomain-containing transcriptional regulators that operate differential genetic programs along the anterior-posterior axis of animal bodies [4]. The domain binds DNA through a helix-turn-helix (HTH) structure. The HTH motif is characterised by two alpha-helices, which make intimate contacts with the DNA and are joined by a short turn. The second helix binds to DNA via a number of hydrogen bonds and hydrophobic interactions, which occur between specific side chains and the exposed bases and thymine methyl groups within the major groove of the DNA [3]. The first helix helps to stabilise the structure.

InterPro



which is needed to avoid steric interference of the beta-carbon with the main chain: for cro and repressor proteins the glycine appears to be mandatory, while for many of the homeotic and other DNA-binding proteins the requirement is relaxed.

Structural links [?]	SCOP a.4.1.1, i.92.1.1 CATH 1.10.10.60 PDB/MSD - click here
Database links [?]	PANDIT PF00046 MSDsite PS00027 Blocks IPB001356 PROSITE doc PDOC00027
Taxonomy [?]	

9	Saccharomyces cerevisiae	Unclassified	
183	Fungi	Virus	7
132	Caenorhabditis elegans	Archaea	7
212	Nematoda	Bacteria	34
4706	Metazoa	Cyanobacteria	2
195	Fruit Fly	Synechocystis PCC 6803	
957	Arthropoda	Rice spp.	167
2531	Chordata	Arabidopsis thaliana	156
405	Mouse	Green Plants	627
371	Human	Plastid Group	630
5341	Eukaryota	Other Eukaryotes	5

1enh

70118 IPF1_MESAU

05527 HMIN_DROME

52955 LBX1_MOUSE

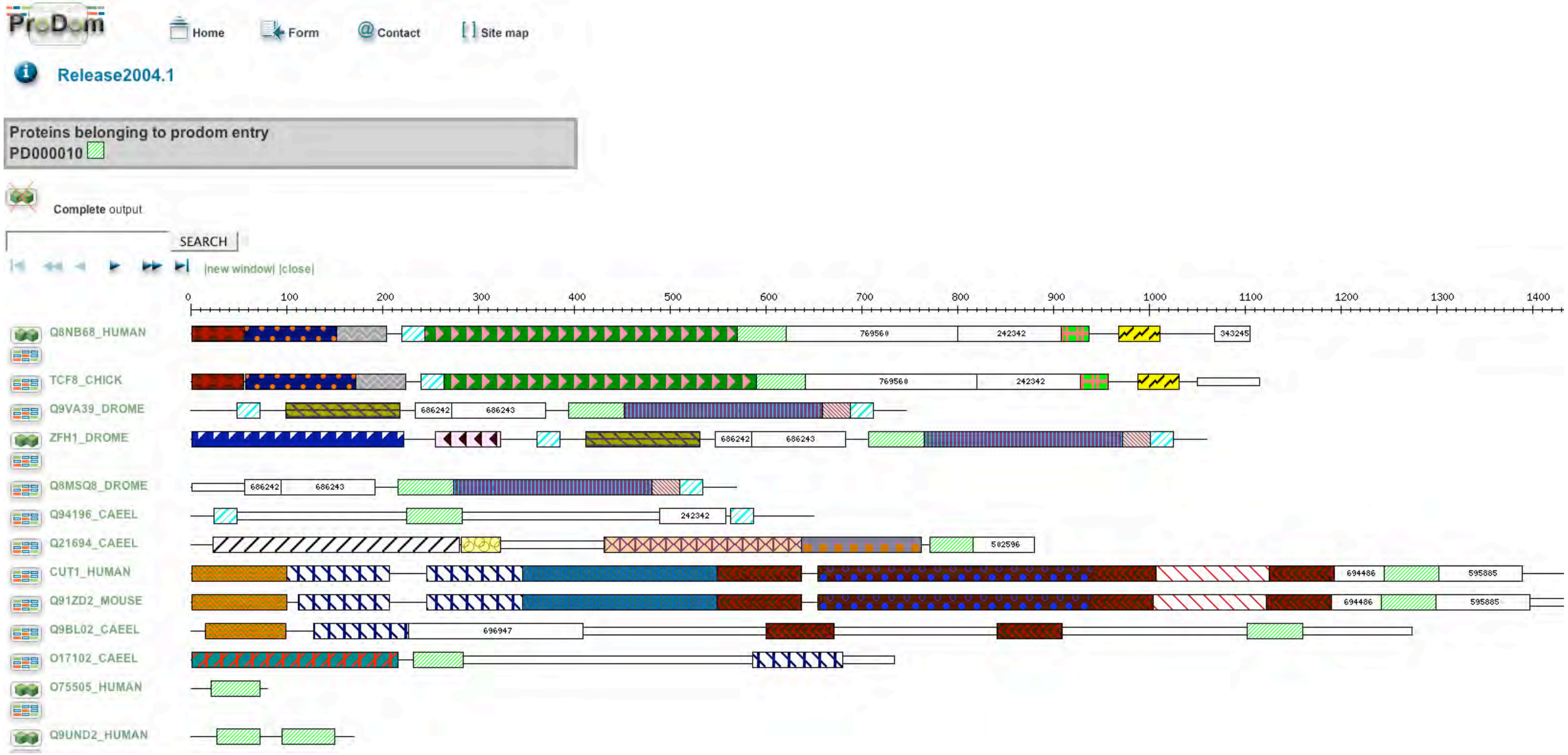
09014 HXB5A_BRARE

41853 RSH1_MAIZE

01366 MATA1_YEAST

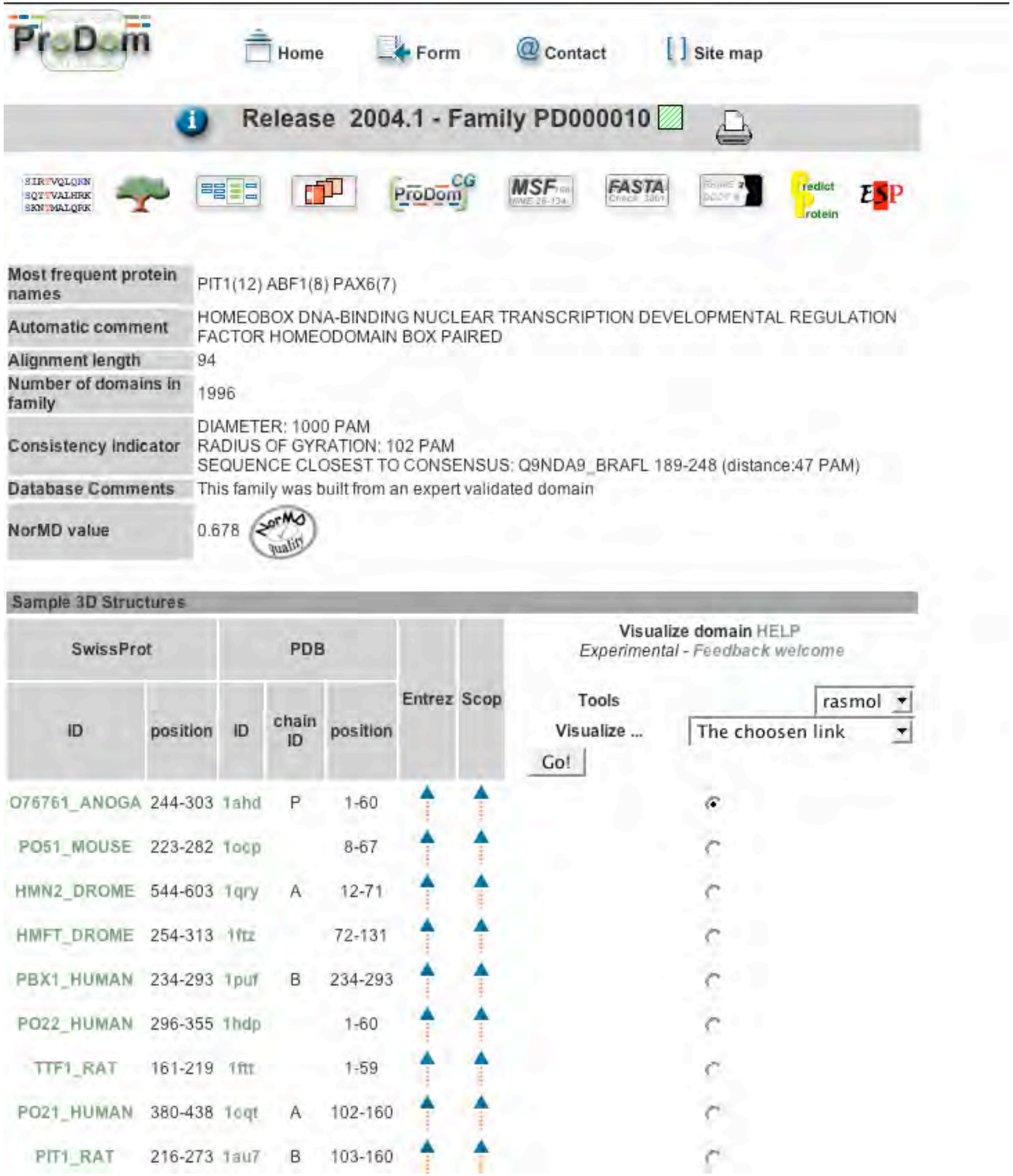
PRODOM

1enh représentation graphique



Proteins belonging to prodom entry PD000010 🟩

PRODOM



Release 2004.1 - Family PD000010 ▨

Most frequent protein names	PIT1(12) ABF1(8) PAX6(7)
Automatic comment	HOMEOBOX DNA-BINDING NUCLEAR TRANSCRIPTION DEVELOPMENTAL REGULATION FACTOR HOMEODOMAIN BOX PAIRED
Alignment length	94
Number of domains in family	1996
Consistency indicator	DIAMETER: 1000 PAM RADIUS OF GYRATION: 102 PAM SEQUENCE CLOSEST TO CONSENSUS: Q9NDA9_BRAFL 189-248 (distance:47 PAM)
Database Comments	This family was built from an expert validated domain
NorMD value	0.678

Sample 3D Structures

Visualize domain HELP
Experimental - Feedback welcome

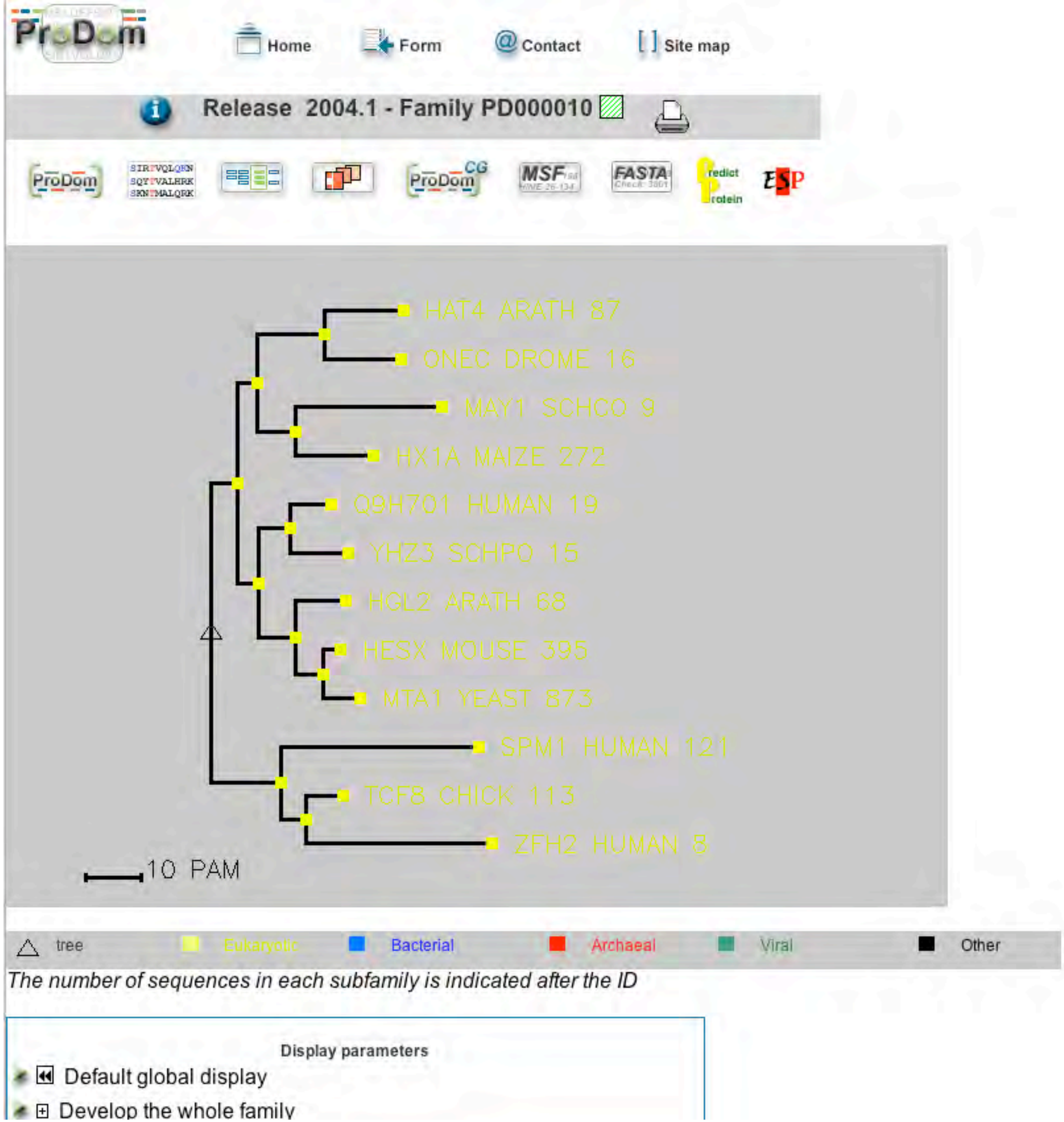
Tools: rasmol
Visualize ...: The choosen link
Go!

SwissProt		PDB			Entrez	Scop	
ID	position	ID	chain ID	position			
O76761_ANOGA	244-303	1ahd	P	1-60	▲	▲	●
PO51_MOUSE	223-282	1ocp		8-67	▲	▲	○
HMN2_DROME	544-603	1qry	A	12-71	▲	▲	○
HMFT_DROME	254-313	1ftz		72-131	▲	▲	○
PBX1_HUMAN	234-293	1puf	B	234-293	▲	▲	○
PO22_HUMAN	296-355	1hdp		1-60	▲	▲	○
TTF1_RAT	161-219	1ftt		1-59	▲	▲	○
PO21_HUMAN	380-438	1oqt	A	102-160	▲	▲	○
PIT1_RAT	216-273	1au7	B	103-160	▲	▲	○

1enh alignement
De la famille

PRODOM

1enh arbre phylogénique



The number of sequences in each subfamily is indicated after the ID

Display parameters

- Default global display
- Develop the whole family

SMART

1enh

Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857-5864
Letunic et al. (2004) *Nucleic Acids Res* **32**, D142-D144

HOX

Homeodomain

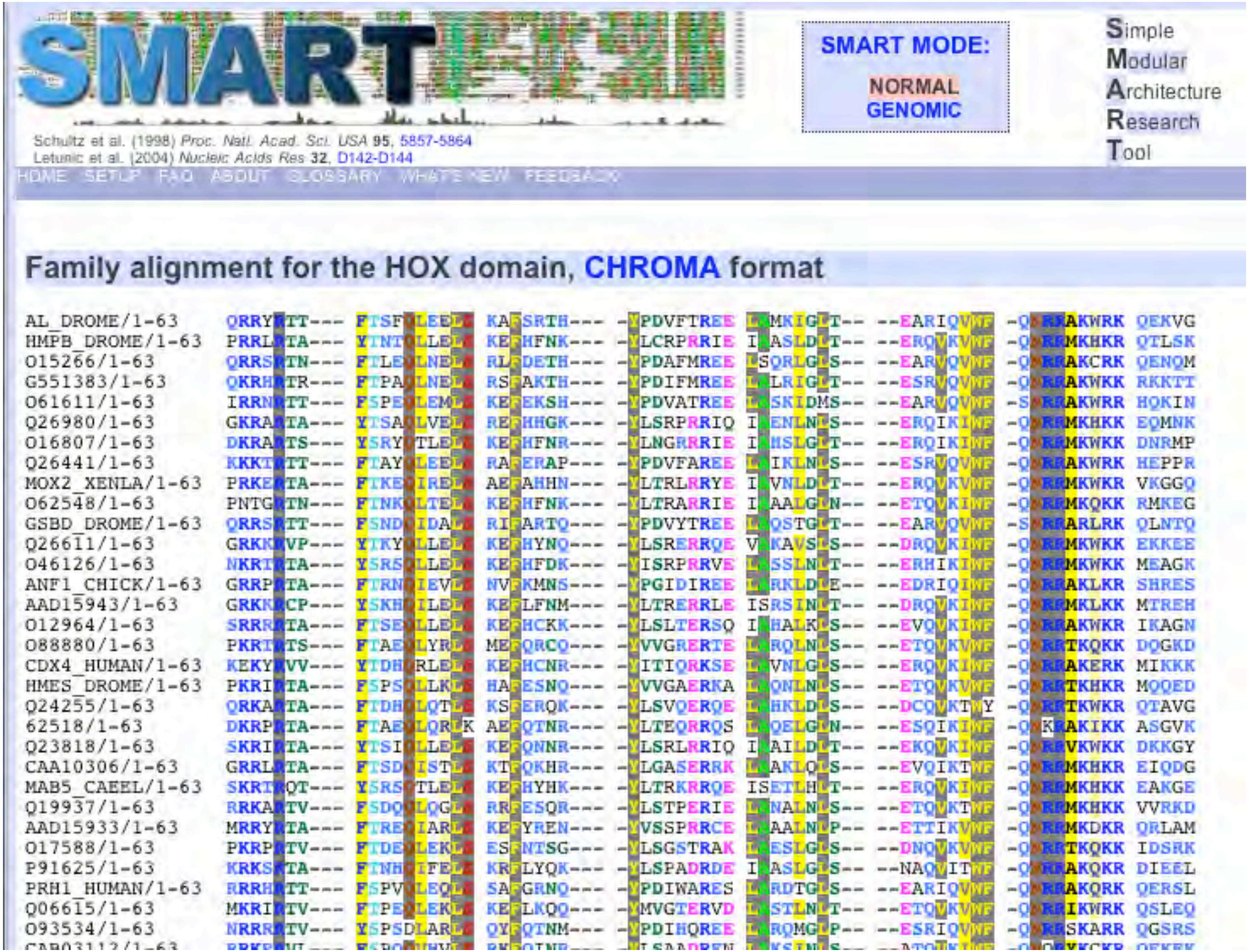
SMART accession number:	SM00389
Description:	DNA-binding factors that are involved in the transcriptional regulation of key developmental processes
Interpro abstract (IPR001356):	The homeobox domain was first identified in a number of drosophila homeotic and segmentation proteins, but is now known to be well-conserved in many other animals, including vertebrates. Hox genes encode homeodomain-containing transcriptional regulators that operate differential genetic programs along the anterior-posterior axis of animal bodies. The domain binds DNA through a helix-turn-helix (HTH) structure. The HTH motif is characterised by two alpha-helices, which make intimate contacts with the DNA and are joined by a short turn. The second helix binds to DNA via a number of hydrogen bonds and hydrophobic interactions, which occur between specific side chains and the exposed bases and thymine methyl groups within the major groove of the DNA. The first helix helps to stabilise the structure. The motif is very similar in sequence and structure in a wide range of DNA-binding proteins (e.g., cro and repressor proteins, homeotic proteins, etc.). One of the principal differences between HTH motifs in these different proteins arises from the stereo-chemical requirement for glycine in the turn which is needed to avoid steric interference of the beta-carbon with the main chain: for cro and repressor proteins the glycine appears to be mandatory, while for many of the homeotic and other DNA-binding proteins the requirement is relaxed.
GO process:	regulation of transcription, DNA-dependent (GO:0006355)
GO component:	nucleus (GO:0005634)
GO function:	transcription factor activity (GO:0003700), DNA binding (GO:0003677)
Family alignment:	View Alignment consensus sequence or Family alignment in CHROMA format

There are 5365 HOX domains in 5125 proteins in SMART's nrdb database.

Click on the following links for more information.

- **Evolution (species in which this domain is found)**
- **Cellular role (predicted cellular role)**
- **Literature (relevant references for this domain)**
- **Disease (disease genes where sequence variants are found in this domain)**
- **Structure (3D structures containing this domain)**
- **Links (links to other resources describing this domain)**

Design and © Ivica Letunic, 2004-2005. Valid XHTML Strict 1.0 and CSS2. Best viewed with a standards compliant browser, like Firefox or Opera.



Simple
Modular
Architecture
Research
Tool

Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857-5864
Letunic et al. (2004) *Nucleic Acids Res* **32**, D142-D144

Family alignment for the HOX domain, CHROMA format

```
AL_DROME/1-63     QRRYRTT--- FTSFQLEELE KAFSRTH--- -YPDVFTREE LAMKIGLT-- --EARIQVWF -QNRRAKWRK QEKVG
HMPB_DROME/1-63   PRRLRTA--- YTNTQLLELE KEFHFNK--- -YLCRPRRIE IAASLDLT-- --ERQVKVWF -QNRRMKHKR QTLSK
O15266/1-63       QRRSRTN--- FTLEQLNELE RLFDETH--- -YPDAFMREE LSQRLGLS-- --EARVQVWF -QNRRAKCRK QENQM
G551383/1-63      QKRHRTR--- FTPAQLNELE RSFAKTH--- -YPDIFMREE LALRIGLT-- --ESRVQVWF -QNRRAKWKK RKKTT
O61611/1-63       IRRNRTT--- FSPEQLEMLE KEFEKSH--- -YPDVATREE LASKIDMS-- --EARVQVWF -SNRRAKWRR HQKIN
Q26980/1-63       GKRARTA--- YTSAQLVELE REFHHGK--- -YLSRPRRIQ IAENLNLS-- --ERQIKIWF -QNRRMKHKK EQMNK
O16807/1-63       DKRARTS--- YSRYQTLELE KEFHFNR--- -YLNGRRRIE IAHSLGLT-- --ERQIKIWF -QNRRMKWKK DNRMP
Q26441/1-63       KKKTRTT--- FTAYQLEELE RAFERAP--- -YPDVFAREE LAIKLNLS-- --ESRVQVWF -QNRRAKWRK HEPPR
MOX2_XENLA/1-63   PRKERTA--- FTKEQIRELE AEFAHHN--- -YLTRLRRYE IAVNLDLT-- --ERQVKVWF -QNRRMKWKR VKGGQ
O62548/1-63       PNTGRTN--- FTNKQLTELE KEFHFNK--- -YLTRARRIE IAAALGLN-- --ETQVKIWF -QNRRMKQKK RMKEG
GSBD_DROME/1-63   QRRSRTT--- FSNDQIDALE RIFARTQ--- -YPDVYTREE LAQSTGLT-- --EARVQVWF -SNRRARLRK QLNTQ
Q26611/1-63       GRKKRVP--- YTKYQLLELE KEFHYNQ--- -YLSRERRQE VAKAVSLS-- --DRQVKIWF -QNRRMKWKK EKKEE
O46126/1-63       NKRTRTA--- YSRSQLLELE KEFHFDK--- -YISRPRRVE LASSLNLT-- --ERHIKIWF -QNRRMKWKK MEAGK
ANF1_CHICK/1-63   GRRPRTA--- FTRNQIEVLE NVFKMNS--- -YPGIDIREE LARKLDLE-- --EDRIQIWF -QNRRAKLKR SHRES
AAD15943/1-63     GRKKRCP--- YSKHQILELE KEFLFNM--- -YLTRERRLE ISRSINLT-- --DRQVKIWF -QNRRMKLKK MTREH
O12964/1-63       SRRRRTA--- FTSEQLLELE KEFHCKK--- -YLSLTERSQ IAHALKLS-- --EVQVKIWF -QNRRAKWKR IKAGN
O88880/1-63       PKRTRTS--- FTAEQLYRLE MEFQRCQ--- -YVVGRERTE LARQLNLS-- --ETQVKVWF -QNRRTKQKK DQGKD
CDX4_HUMAN/1-63   KEKYRVV--- YTDHQRLELE KEFHCNR--- -YITIQRKSE LAVNLGLS-- --ERQVKIWF -QNRRAKERK MIKKK
HMES_DROME/1-63   PKRIRTA--- FSPSQLLKLE HAFESNQ--- -YVVGAERKA LAQNLNLS-- --ETQVKVWF -QNRRTKHKR MQQED
Q24255/1-63       QRKARTA--- FTDHQLQTLE KSFERQK--- -YLSVQERQE LAHKLDLS-- --DCQVKTWY -QNRRTKWKR QTAVG
62518/1-63        DKRPRTA--- FTAEQLQRLK AEFQTNR--- -YLTEQRRQS LAQELGLN-- --ESQIKIWF -QNKRAKIKK ASGVK
Q23818/1-63       SKRIRTA--- YTSIQLLELE KEFQNNR--- -YLSRLRRIQ IAAILDLT-- --EKQVKIWF -QNRRVKWKK DKKGY
CAA10306/1-63     GRRLRTA--- FTSDQISTLE KTFQKHR--- -YLGASERRK LAAKLQLS-- --EVQIKTWF -QNRRMKHKR EIQDG
MAB5_CAEEL/1-63   SKRTRQT--- YSRSQTLELE KEFHYHK--- -YLTRKRRQE ISETLHLT-- --ERQVKIWF -QNRRMKHKK EAKGE
Q19937/1-63       RRKARTV--- FSDQQLQGLE RRFESQR--- -YLSTPERIE LANALNLS-- --ETQVKTWF -QNRRMKHKK VVRKD
AAD15933/1-63     MRRYRTA--- FTREQIARLE KEFYREN--- -YVSSPRRCE LAAALNLP-- --ETTIKVWF -QNRRMKDKR QRLAM
O17588/1-63       PKRPRTV--- FTDEQLEKLE ESFNTSG--- -YLSGSTRAK LAESLGLS-- --DNQVKVWF -QNRRTKQKK IDSRK
P91625/1-63       KRKSRTA--- FTNHQIFELE KRFLYQK--- -YLSPADRDE IAASLGLS-- --NAQVITWF -QNRRAKQKR DIEEL
PRH1_HUMAN/1-63   RRRHRTT--- FSPVQLEQLE SAFGRNQ--- -YPDIWARES LARDTGLS-- --EARIQVWF -QNRRAKQRK QERSL
Q06615/1-63       MKRIRTV--- FTPEQLEKLE KEFLKQQ--- -YMVGTERVD LASTLNLT-- --ETQVKVWF -QNRRIKWRK QSLEQ
O93534/1-63       NRRRRTV--- YSPSDLARLE QYFQTNM--- -YPDIHQREE LARQMGLP-- --ESRIQVWF -QNRRSKARR QGSRS
```

SMART

1enh

There are 5365 HOX domains in 5125 proteins in SMART's nrdb database.

Click on the following links for more information.

Evolution (species in which this domain is found)

Taxonomic distribution of proteins with HOX domain.

- Bacteria (3 / 0.06%)
  - undefined (3 / 0.06%)
- Eukaryota (5071 / 98.95%)
  - Fungi (163 / 3.18%)
  - Metazoa (4386 / 85.58%)
  - Viridiplantae (517 / 10.09%)
  - undefined (5 / 0.10%)
- undefined (51 / 1.00%)
  - undefined (51 / 1.00%)

Click on ⊞ to expand nodes. To display all proteins with a HOX domain in a specific node, click on it.

This tree shows only several representative species. The complete taxonomic breakdown of all proteins with HOX domain is also avaliable.

Useful shortcuts: Expand all nodes or Collapse tree
Go to specific node: Anopheles gambiae, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae, Takifugu rubripes

Cellular role (predicted cellular role)

Cellular role: transcription
Binding / catalysis: DNA-binding

Literature (relevant references for this domain)

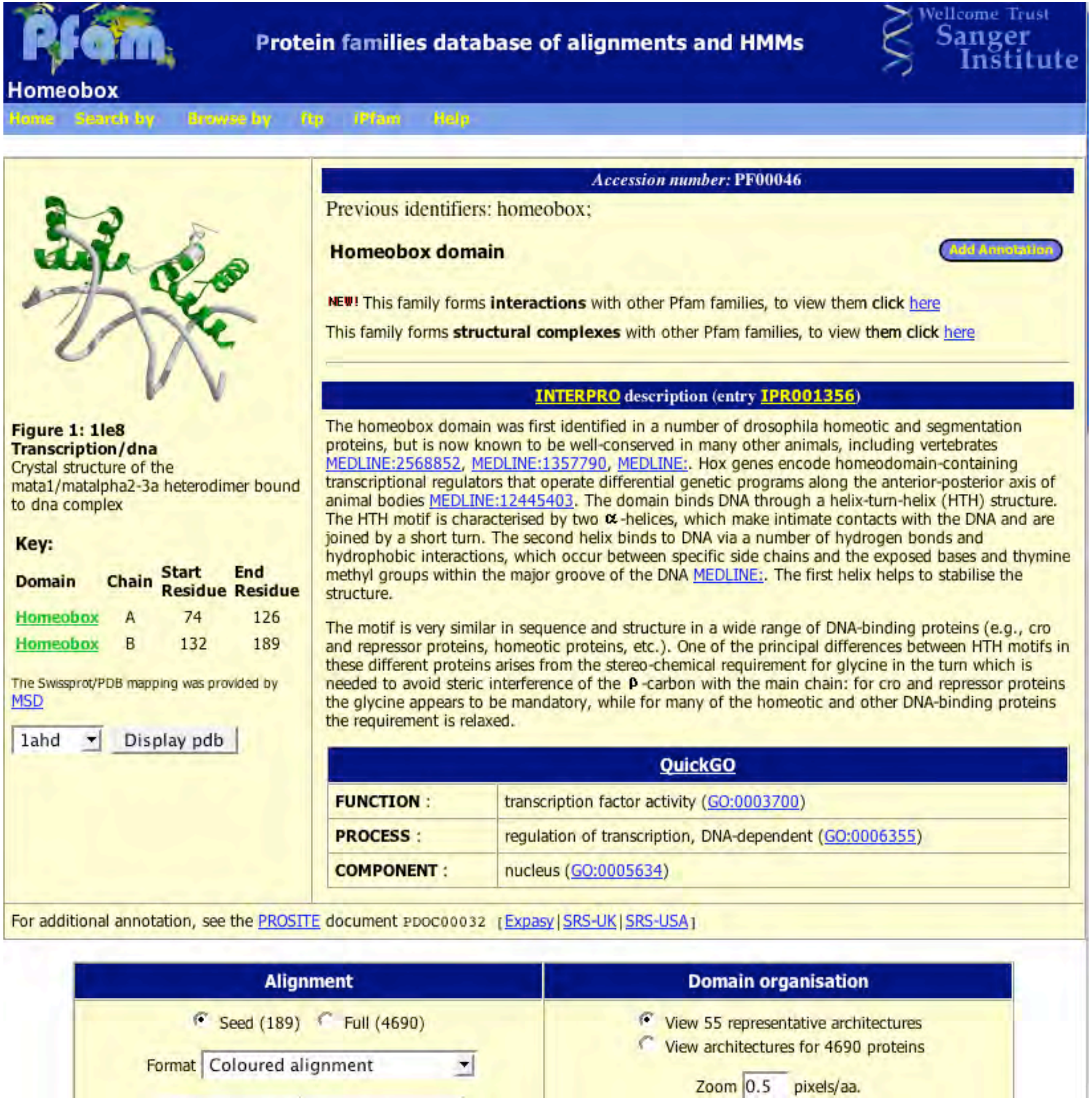
Disease (disease genes where sequence variants are found in this domain)

SwissProt sequences and OMIM curated human diseases associated with missense mutations within the HOX domain.

Protein	Disease
undefined function (SRS)(SMART)	OMIM:600584: Atrial septal defect with atrioventricular conduction defects OMIM:108900:
Homeobox protein MSX-1 (Hox-7). (SRS)(SMART)	OMIM:142983: Hypodontia, autosomal dominant OMIM:106600: Hypodontia with orofacial cleft OMIM:106600:
Homeobox expressed in ES cells 1 (Homeobox protein ANF) (hAnf). (SRS)(SMART)	OMIM:601802: Septooptic dysplasia OMIM:182230:
Pituitary-specific positive transcription factor 1 (Pit-1) (Growth hormone factor 1) (GHF-1). (SRS)(SMART)	OMIM:173110: Pituitary hormone deficiency, combined
Pituitary homeobox 2 (RIEG bicoid-related homeobox transcription factor) (Solurshin) (ALL1 responsive protein ARP1). (SRS)(SMART)	OMIM:137600: Iridogoniodysgenesis syndrome OMIM:601542: Rieger syndrome OMIM:180500: Iridogoniodysgenesis syndrome-2 OMIM:137600:
	OMIM:168500: Parietal foramina

PFAM

1enh



Accession number: **PF00046**

Previous identifiers: homeobox;

Homeobox domain

NEW! This family forms **interactions** with other Pfam families, to view them click here

This family forms **structural complexes** with other Pfam families, to view them click here

INTERPRO description (entry IPR001356)

The homeobox domain was first identified in a number of drosophila homeotic and segmentation proteins, but is now known to be well-conserved in many other animals, including vertebrates MEDLINE:2568852, MEDLINE:1357790, MEDLINE:. Hox genes encode homeodomain-containing transcriptional regulators that operate differential genetic programs along the anterior-posterior axis of animal bodies MEDLINE:12445403. The domain binds DNA through a helix-turn-helix (HTH) structure. The HTH motif is characterised by two α-helices, which make intimate contacts with the DNA and are joined by a short turn. The second helix binds to DNA via a number of hydrogen bonds and hydrophobic interactions, which occur between specific side chains and the exposed bases and thymine methyl groups within the major groove of the DNA MEDLINE:. The first helix helps to stabilise the structure.

The motif is very similar in sequence and structure in a wide range of DNA-binding proteins (e.g., cro and repressor proteins, homeotic proteins, etc.). One of the principal differences between HTH motifs in these different proteins arises from the stereo-chemical requirement for glycine in the turn which is needed to avoid steric interference of the β-carbon with the main chain: for cro and repressor proteins the glycine appears to be mandatory, while for many of the homeotic and other DNA-binding proteins the requirement is relaxed.

Figure 1: 1le8 Transcription/dna
Crystal structure of the mata1/matalpha2-3a heterodimer bound to dna complex

Key:

Domain	**Chain**	**Start Residue**	**End Residue**
Homeobox	A	74	126
Homeobox	B	132	189

The Swissprot/PDB mapping was provided by MSD

QuickGO	
FUNCTION :	transcription factor activity (GO:0003700)
PROCESS :	regulation of transcription, DNA-dependent (GO:0006355)
COMPONENT :	nucleus (GO:0005634)

For additional annotation, see the PROSITE document PDOC00032 [Expasy | SRS-UK | SRS-USA]

Alignment

● Seed (189) ○ Full (4690)

Format: Coloured alignment

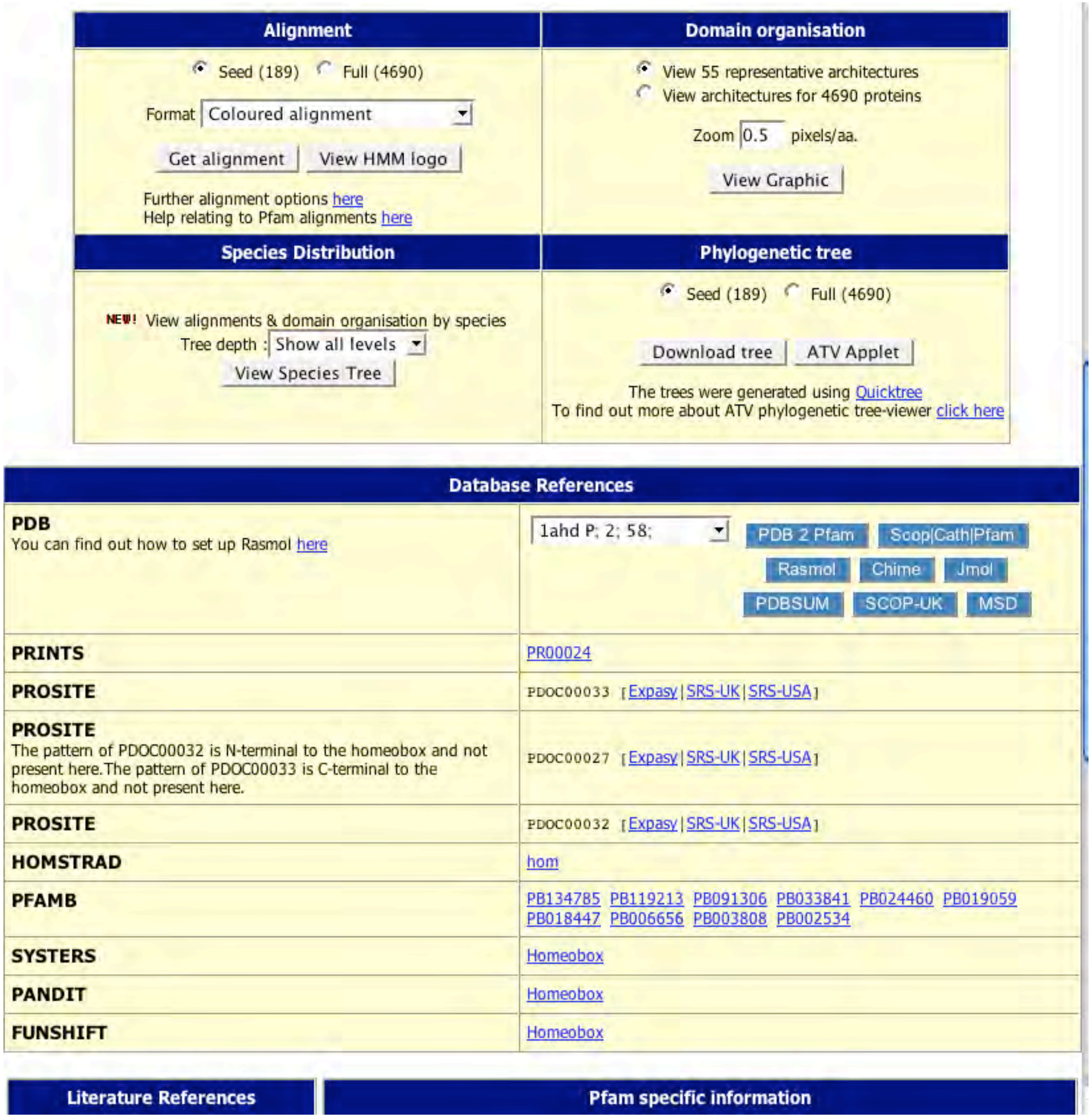
Domain organisation

● View 55 representative architectures
○ View architectures for 4690 proteins

Zoom 0.5 pixels/aa.

PFAM

1enh



Format: Coloured alignment

Further alignment options here
Help relating to Pfam alignments here

● View 55 representative architectures
○ View architectures for 4690 proteins

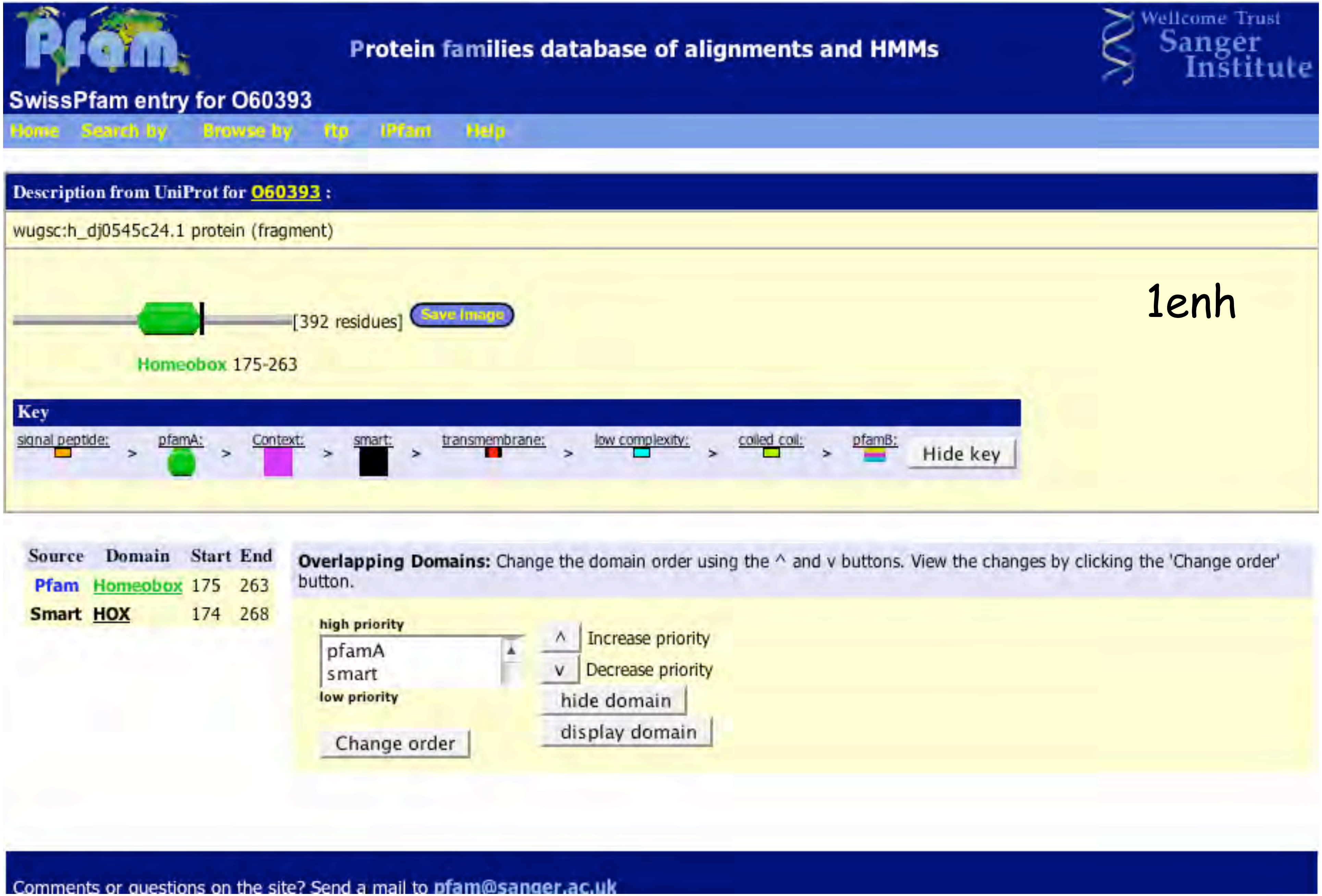
Zoom: 0.5 pixels/aa.

View alignments & domain organisation by species
Tree depth : Show all levels

The trees were generated using Quicktree
To find out more about ATV phylogenetic tree-viewer click here

PDB You can find out how to set up Rasmol here	1ahd P; 2; 58;
PRINTS	PR00024
PROSITE	PDOC00033 [Expasy | SRS-UK | SRS-USA]
PROSITE The pattern of PDOC00032 is N-terminal to the homeobox and not present here. The pattern of PDOC00033 is C-terminal to the homeobox and not present here.	PDOC00027 [Expasy | SRS-UK | SRS-USA]
PROSITE	PDOC00032 [Expasy | SRS-UK | SRS-USA]
HOMSTRAD	hom
PFAMB	PB134785 PB119213 PB091306 PB033841 PB024460 PB019059 PB018447 PB006656 PB003808 PB002534
SYSTERS	Homeobox
PANDIT	Homeobox
FUNSHIFT	Homeobox

PFAM



SwissPfam entry for O60393

Description from UniProt for O60393 :

wugsc:h_dj0545c24.1 protein (fragment)

[392 residues]

Homeobox 175-263

1enh

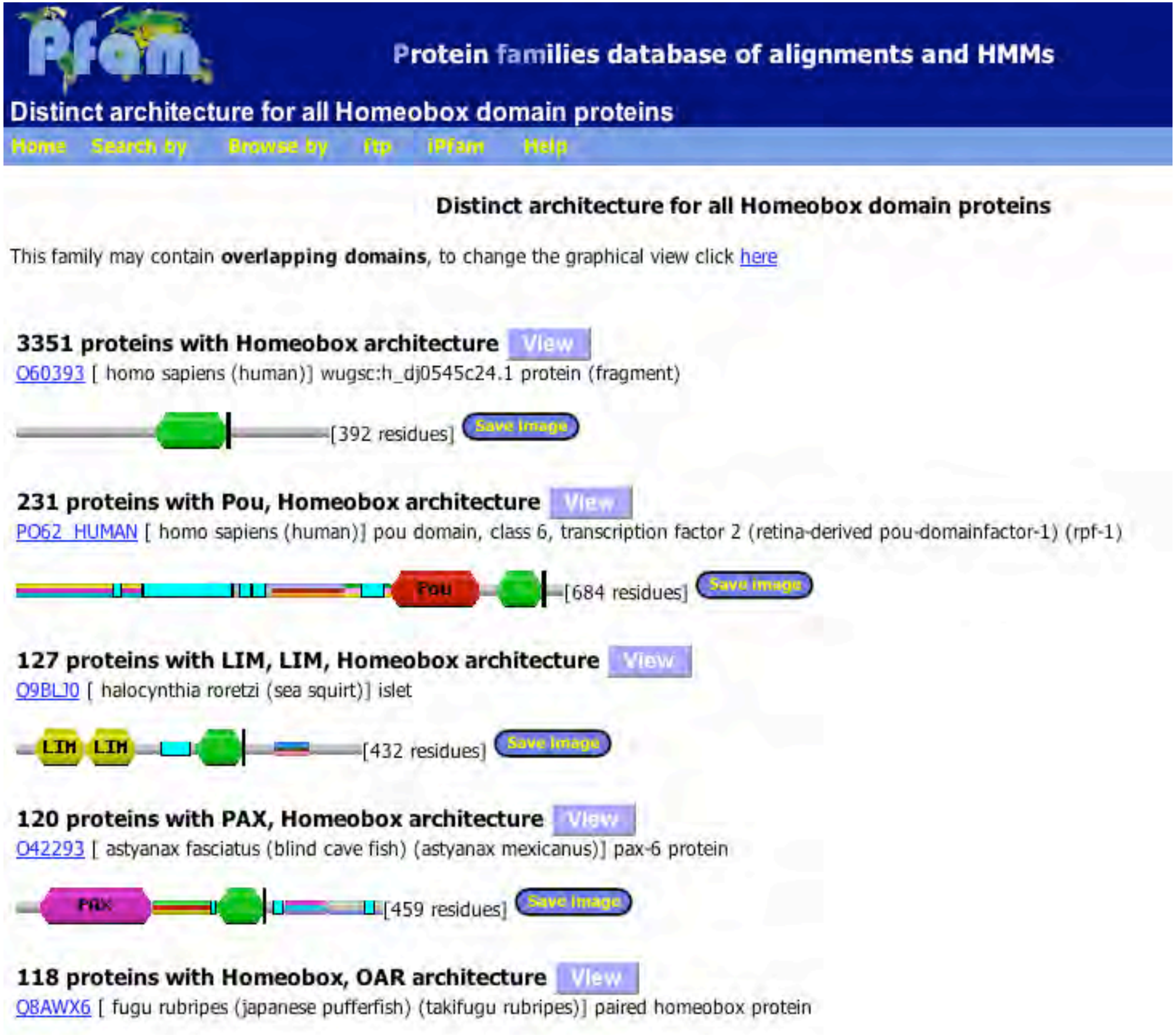
Source	Domain	Start	End
Pfam	Homeobox	175	263
Smart	HOX	174	268

Overlapping Domains: Change the domain order using the ^ and v buttons. View the changes by clicking the 'Change order' button.

Comments or questions on the site? Send a mail to pfam@sanger.ac.uk

PFAM

1enh



Protein families database of alignments and HMMs

Distinct architecture for all Homeobox domain proteins

Distinct architecture for all Homeobox domain proteins

This family may contain **overlapping domains**, to change the graphical view click here

3351 proteins with Homeobox architecture
O60393 [homo sapiens (human)] wugsc:h_dj0545c24.1 protein (fragment)

[392 residues]

231 proteins with Pou, Homeobox architecture
PO62_HUMAN [homo sapiens (human)] pou domain, class 6, transcription factor 2 (retina-derived pou-domainfactor-1) (rpf-1)

[684 residues]

127 proteins with LIM, LIM, Homeobox architecture
Q9BLJ0 [halocynthia roretzi (sea squirt)] islet

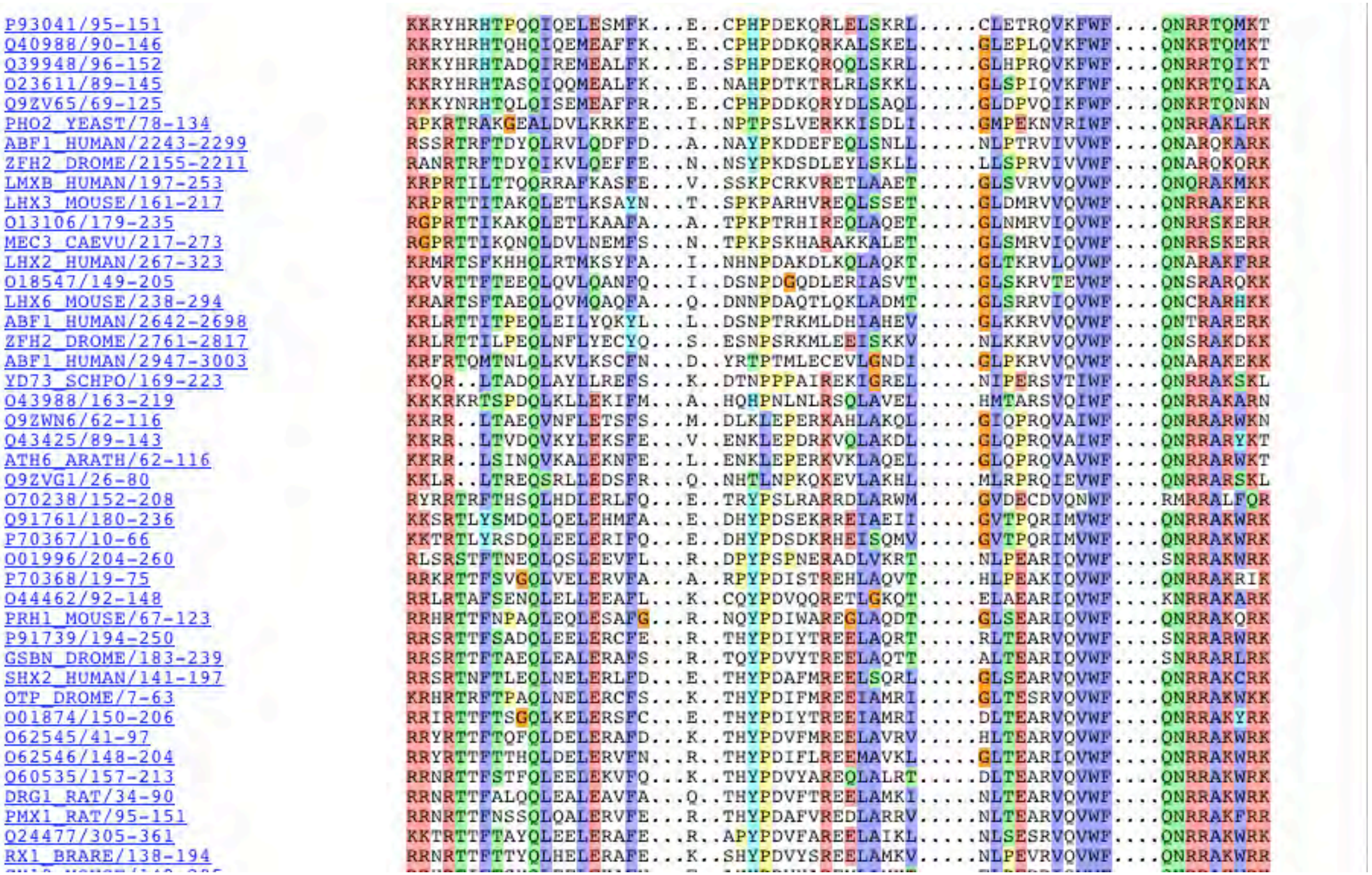
[432 residues]

120 proteins with PAX, Homeobox architecture
O42293 [astyanax fasciatus (blind cave fish) (astyanax mexicanus)] pax-6 protein

[459 residues]

118 proteins with Homeobox, OAR architecture
Q8AWX6 [fugu rubripes (japanese pufferfish) (takifugu rubripes)] paired homeobox protein

PFAM

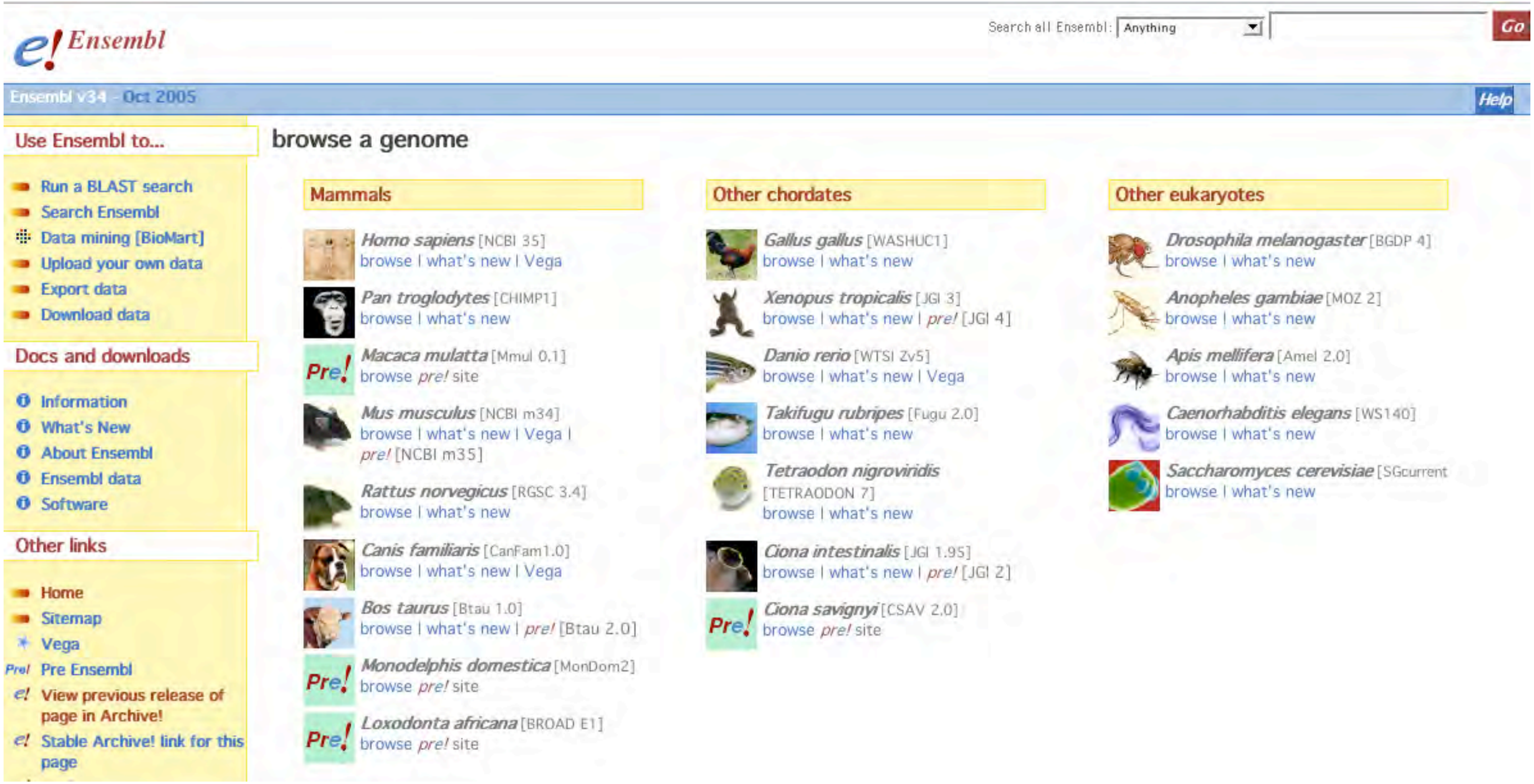


```
P93041/95-151          KKRYHRHTPQQIQELESMFK...E..CPHPDEKQRLELSKRL.....CLETRQVKFWF....QNRRTQMKT
Q40988/90-146          KKRYHRHTQHQIQEMEAFFK...E..CPHPDDKQRKALSKEL.....GLEPLQVKFWF....QNKRTQMKT
Q39948/96-152          RKKYHRHTADQIREMEALFK...E..SPHPDEKQRQQLSKRL.....GLHPRQVKFWF....QNRRTQIKT
O23611/89-145          KKRYHRHTASQIQQMEALFK...E..NAHPDTKTRLRLSKKL.....GLSPIQVKFWF....QNKRTQIKA
Q9ZV65/69-125          KKKYNRHTQLQISEMEAFFR...E..CPHPDDKQRYDLSAQL.....GLDPVQIKFWF....QNKRTQNKN
PHO2_YEAST/78-134      RPKRTRAKGEALDVLKRKFE...I..NPTPSLVERKKISDLI.....GMPEKNVRIWF....QNRRAKLRK
ABF1_HUMAN/2243-2299   RSSRTRFTDYQLRVLQDFFD...A..NAYPKDDEFEQLSNLL.....NLPTRVIVVWF....QNARQKARK
ZFH2_DROME/2155-2211   RANRTRFTDYQIKVLQEFFE...N..NSYPKDSDLEYLSKLL.....LLSPRVIVVWF....QNARQKQRK
LMXB_HUMAN/197-253     KRPRTILTTQQRRAFKASFE...V..SSKPCRKVRETLAAET.....GLSVRVVQVWF....QNQRAKMKK
LHX3_MOUSE/161-217     KRPRTTITAKQLETLKSAYN...T..SPKPARHVREQLSSET.....GLDMRVVQVWF....QNRRAKEKR
O13106/179-235         RGPRTTIKAKQLETLKAAFA...A..TPKPTRHIREQLAQET.....GLNMRVIQVWF....QNRRSKERR
MEC3_CAEVU/217-273     RGPRTTIKQNQLDVLNEMFS...N..TPKPSKHARAKKALET.....GLSMRVIQVWF....QNRRSKERR
LHX2_HUMAN/267-323     KRMRTSFKHHQLRTMKSYFA...I..NHNPDAKDLKQLAQKT.....GLTKRVLQVWF....QNARAKFRR
O18547/149-205         KRVRTTFTEEQLQVLQANFQ...I..DSNPDGQDLERIASVT.....GLSKRVTEVWF....QNSRARQKK
LHX6_MOUSE/238-294     KRARTSFTAEQLQVMQAQFA...Q..DNNPDAQTLQKLADMT.....GLSRRVIQVWF....QNCRARHKK
ABF1_HUMAN/2642-2698   KRLRTTITPEQLEILYQKYL...L..DSNPTRKMLDHIAHEV.....GLKKRVVQVWF....QNTRARERK
ZFH2_DROME/2761-2817   KRLRTTILPEQLNFLYECYQ...S..ESNPSRKMLEEISKKV.....NLKKRVVQVWF....QNSRAKDKK
ABF1_HUMAN/2947-3003   KRFRTQMTNLQLKVLKSCFN...D..YRTPTMLECEVLGNDI.....GLPKRVVQVWF....QNARAKEKK
YD73_SCHPO/169-223     KKQR..LTADQLAYLLREFS...K..DTNPPPAIREKIGREL.....NIPERSVTIWF....QNRRAKSKL
O43988/163-219         KKKRKRTSPDQLKLLEKIFM...A..HQHPNLNLRSQLAVEL.....HMTARSVQIWF....QNRRAKARN
Q9ZWN6/62-116          KKRR..LTAEQVNFLETSFS...M..DLKLEPERKAHLAKQL.....GIQPRQVAIWF....QNRRARWKN
Q43425/89-143          KKRR..LTVDQVKYLEKSFE...V..ENKLEPDRKVQLAKDL.....GLQPRQVAIWF....QNRRARYKT
ATH6_ARATH/62-116      KKRR..LSINQVKALEKNFE...L..ENKLEPERKVKLAQEL.....GLQPRQVAVWF....QNRRARWKT
Q9ZVG1/26-80           KKLR..LTREQSRLLEDSFR...Q..NHTLNPKQKEVLAKHL.....MLRPRQIEVWF....QNRRARSKL
O70238/152-208         RYRRTRFTHSQLHDLERLFQ...E..TRYPSLRARRDLARWM.....GVDECDVQNWF....RMRRALFQR
Q91761/180-236         KKSRTLYSMDQLQELEHMFA...E..DHYPDSEKRREIAEII.....GVTPQRIMVWF....QNRRAKWRK
P70367/10-66           KKTRTLYRSDQLEELERIFQ...E..DHYPDSDKRHEISQMV.....GVTPQRIMVWF....QNRRAKWRK
O01996/204-260         RLSRSTFTNEQLQSLEEVFL...R..DPYPSPNERADLVKRT.....NLPEARIQVWF....SNRRAKWRK
P70368/19-75           RRKRTTFSVGQLVELERVFA...A..RPYPDISTREHLAQVT.....HLPEAKIQVWF....QNRRAKRIK
O44462/92-148          RRLRTAFSENQLELLEEAFL...K..CQYPDVQQRETLGKQT.....ELAEARIQVWF....KNRRAKARK
PRH1_MOUSE/67-123      RRHRTTFNPAQLEQLESAFG...R..NQYPDIWAREGLAQDT.....GLSEARIQVWF....QNRRAKQRK
P91739/194-250         RRSRTTFSADQLEELERCFE...R..THYPDIYTREELAQRT.....RLTEARVQVWF....SNRRARWRK
GSBN_DROME/183-239     RRSRTTFTAEQLEALERAFS...R..TQYPDVYTREELAQTT.....ALTEARIQVWF....SNRRARLRK
SHX2_HUMAN/141-197     RRSRTNFTLEQLNELERLFD...E..THYPDAFMREELSQRL.....GLSEARVQVWF....QNRRAKCRK
OTP_DROME/7-63         KRHRTRFTPAQLNELERCFS...K..THYPDIFMREEIAMRI.....GLTESRVQVWF....QNRRAKWKK
O01874/150-206         RRIRTTFTSGQLKELERSFC...E..THYPDIYTREEIAMRI.....DLTEARVQVWF....QNRRAKYRK
O62545/41-97           RRYRTTFTQFQLDELERAFD...K..THYPDVFMREELAVRV.....HLTEARVQVWF....QNRRAKWRK
O62546/148-204         RRYRTTFTTHQLDELERVFN...R..THYPDIFLREEMAVKL.....GLTEARIQVWF....QNRRAKWRK
O60535/157-213         RRNRTTFSTFQLEELEKVFQ...K..THYPDVYAREQLALRT.....DLTEARVQVWF....QNRRAKWRK
DRG1_RAT/34-90         RRNRTTFALQQLEALEAVFA...Q..THYPDVFTREELAMKI.....NLTEARVQVWF....QNRRAKWRK
PMX1_RAT/95-151        RRNRTTFNSSQLQALERVFE...R..THYPDAFVREDLARRV.....NLTEARVQVWF....QNRRAKFRR
Q24477/305-361         KKTRTTFTAYQLEELERAFE...R..APYPDVFAREELAIKL.....NLSESRVQVWF....QNRRAKWRK
RX1_BRARE/138-194      RRNRTTFTTYQLHELERAFE...K..SHYPDVYSREELAMKV.....NLPEVRVQVWF....QNRRAKWRR
```

1enh

Génomes complets

http://www.ensembl.org/index.html



Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- Stable Archive! link for this page

browse a genome

Mammals

- *Homo sapiens* [NCBI 35] browse | what's new | Vega
- *Pan troglodytes* [CHIMP1] browse | what's new
- *Macaca mulatta* [Mmul 0.1] browse *pre!* site
- *Mus musculus* [NCBI m34] browse | what's new | Vega | *pre!* [NCBI m35]
- *Rattus norvegicus* [RGSC 3.4] browse | what's new
- *Canis familiaris* [CanFam1.0] browse | what's new | Vega
- *Bos taurus* [Btau 1.0] browse | what's new | *pre!* [Btau 2.0]
- *Monodelphis domestica* [MonDom2] browse *pre!* site
- *Loxodonta africana* [BROAD E1] browse *pre!* site

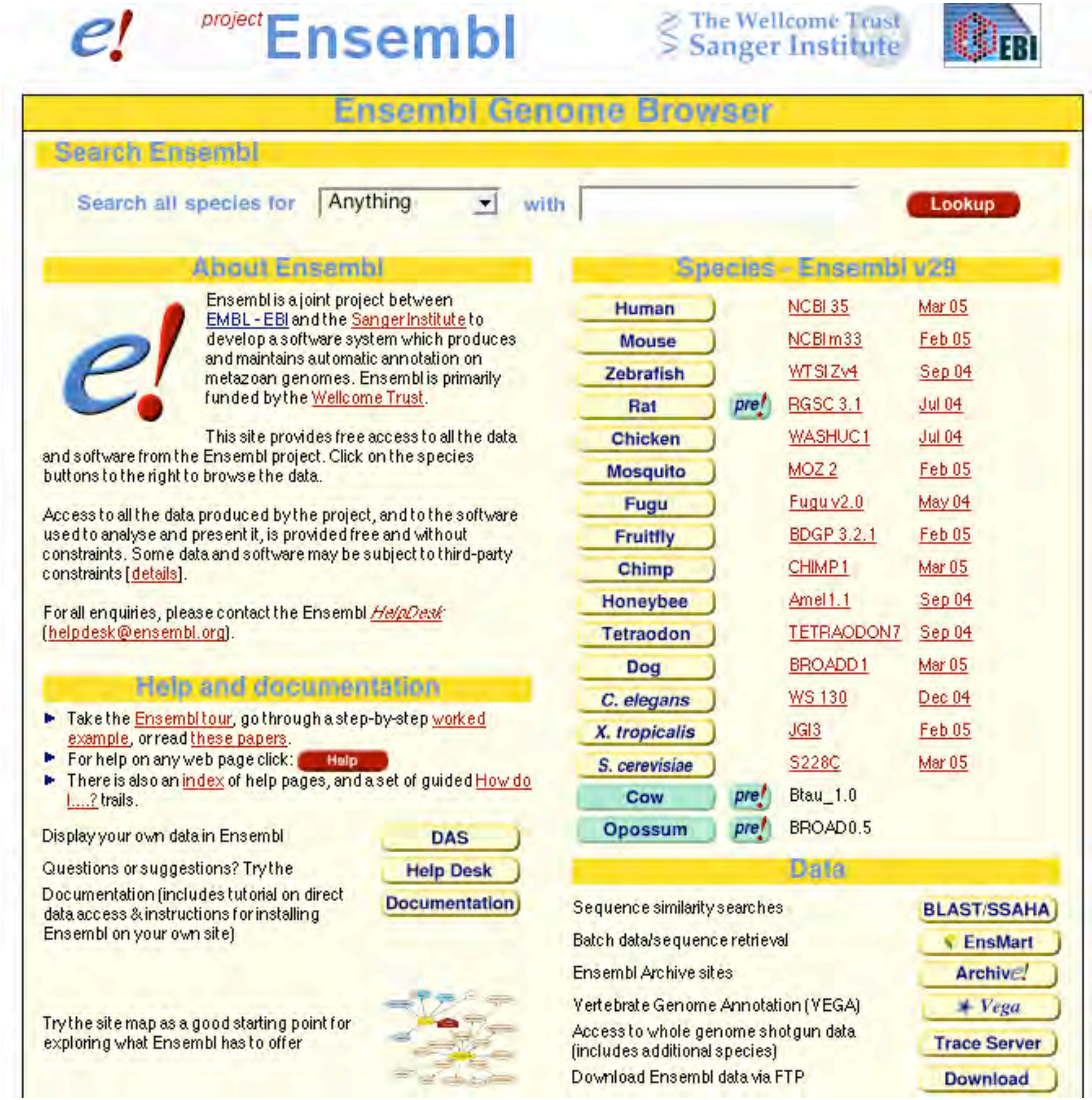
Other chordates

- *Gallus gallus* [WASHUC1] browse | what's new
- *Xenopus tropicalis* [JGI 3] browse | what's new | *pre!* [JGI 4]
- *Danio rerio* [WTSI Zv5] browse | what's new | Vega
- *Takifugu rubripes* [Fugu 2.0] browse | what's new
- *Tetraodon nigroviridis* [TETRAODON 7] browse | what's new
- *Ciona intestinalis* [JGI 1.95] browse | what's new | *pre!* [JGI 2]
- *Ciona savignyi* [CSAV 2.0] browse *pre!* site

Other eukaryotes

- *Drosophila melanogaster* [BGDP 4] browse | what's new
- *Anopheles gambiae* [MOZ 2] browse | what's new
- *Apis mellifera* [Amel 2.0] browse | what's new
- *Caenorhabditis elegans* [WS140] browse | what's new
- *Saccharomyces cerevisiae* [SGcurrent browse | what's new

EnsEMBL



Ensembl Genome Browser

Search Ensembl

Search all species for Anything with

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on metazoan genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on the species buttons to the right to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints [details].

For all enquiries, please contact the Ensembl *HelpDesk* (helpdesk@ensembl.org).

Help and documentation

- Take the Ensembl tour, go through a step-by-step worked example, or read these papers.
- For help on any web page click: Help
- There is also an index of help pages, and a set of guided How do I....? trails.

Display your own data in Ensembl — DAS

Questions or suggestions? Try the — Help Desk

Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site) — Documentation

Try the site map as a good starting point for exploring what Ensembl has to offer

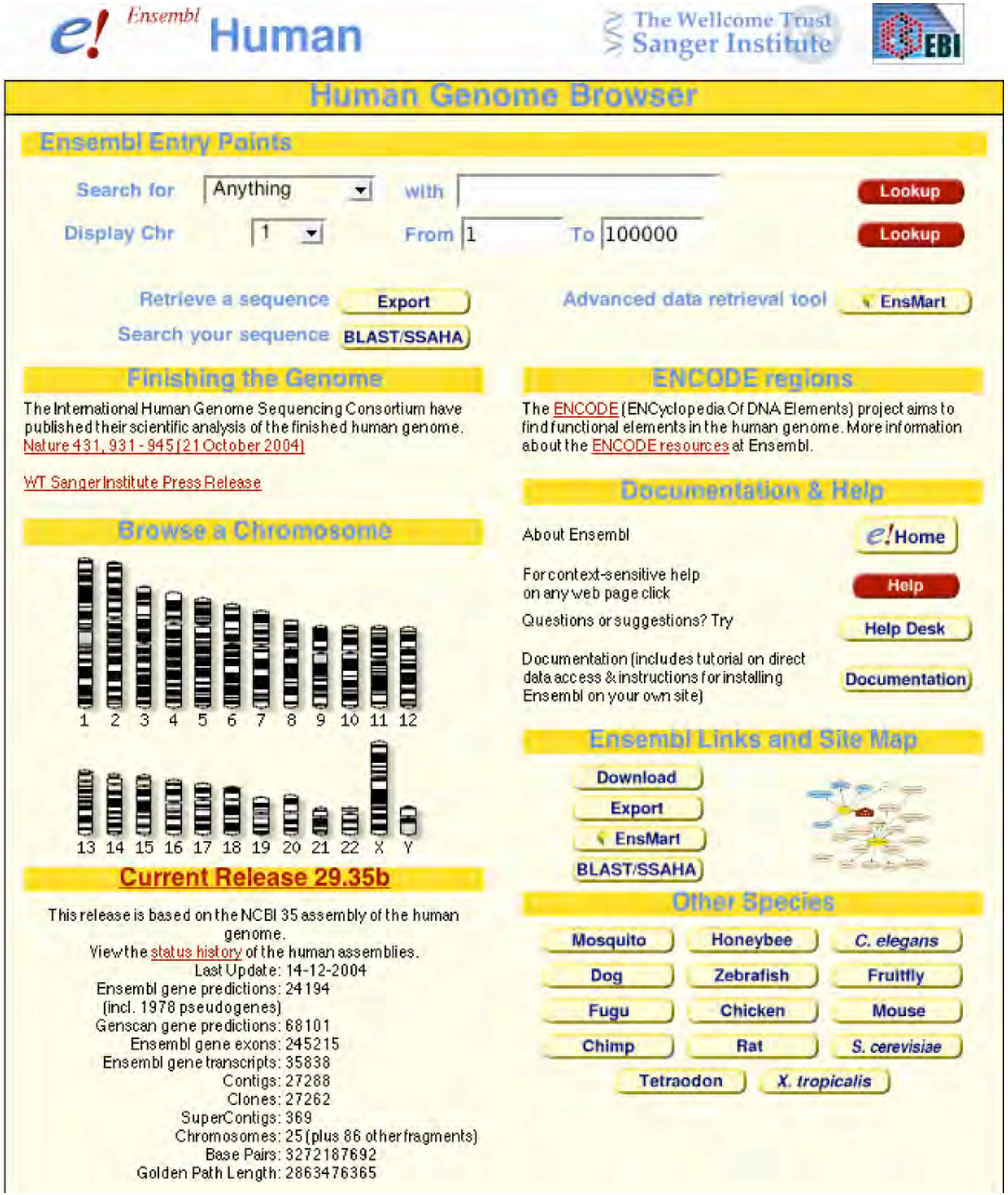
Species - Ensembl v29

Human		NCBI 35	Mar 05
Mouse		NCBI m33	Feb 05
Zebrafish		WTSI Zv4	Sep 04
Rat	pre!	RGSC 3.1	Jul 04
Chicken		WASHUC1	Jul 04
Mosquito		MOZ 2	Feb 05
Fugu		Fugu v2.0	May 04
Fruitfly		BDGP 3.2.1	Feb 05
Chimp		CHIMP1	Mar 05
Honeybee		Amel 1.1	Sep 04
Tetraodon		TETRAODON7	Sep 04
Dog		BROADD1	Mar 05
C. elegans		WS 130	Dec 04
X. tropicalis		JGI3	Feb 05
S. cerevisiae		S228C	Mar 05
Cow	pre!	Btau_1.0	
Opossum	pre!	BROAD0.5	

Data

Sequence similarity searches	BLAST/SSAHA
Batch data/sequence retrieval	EnsMart
Ensembl Archive sites	Archive!
Vertebrate Genome Annotation (VEGA)	Vega
Access to whole genome shotgun data (includes additional species)	Trace Server
Download Ensembl data via FTP	Download

EnsEMBL



The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome. Nature 431, 931 - 945 (21 October 2004)

WT Sanger Institute Press Release

The ENCODE (ENCyclopedia Of DNA Elements) project aims to find functional elements in the human genome. More information about the ENCODE resources at Ensembl.

About Ensembl

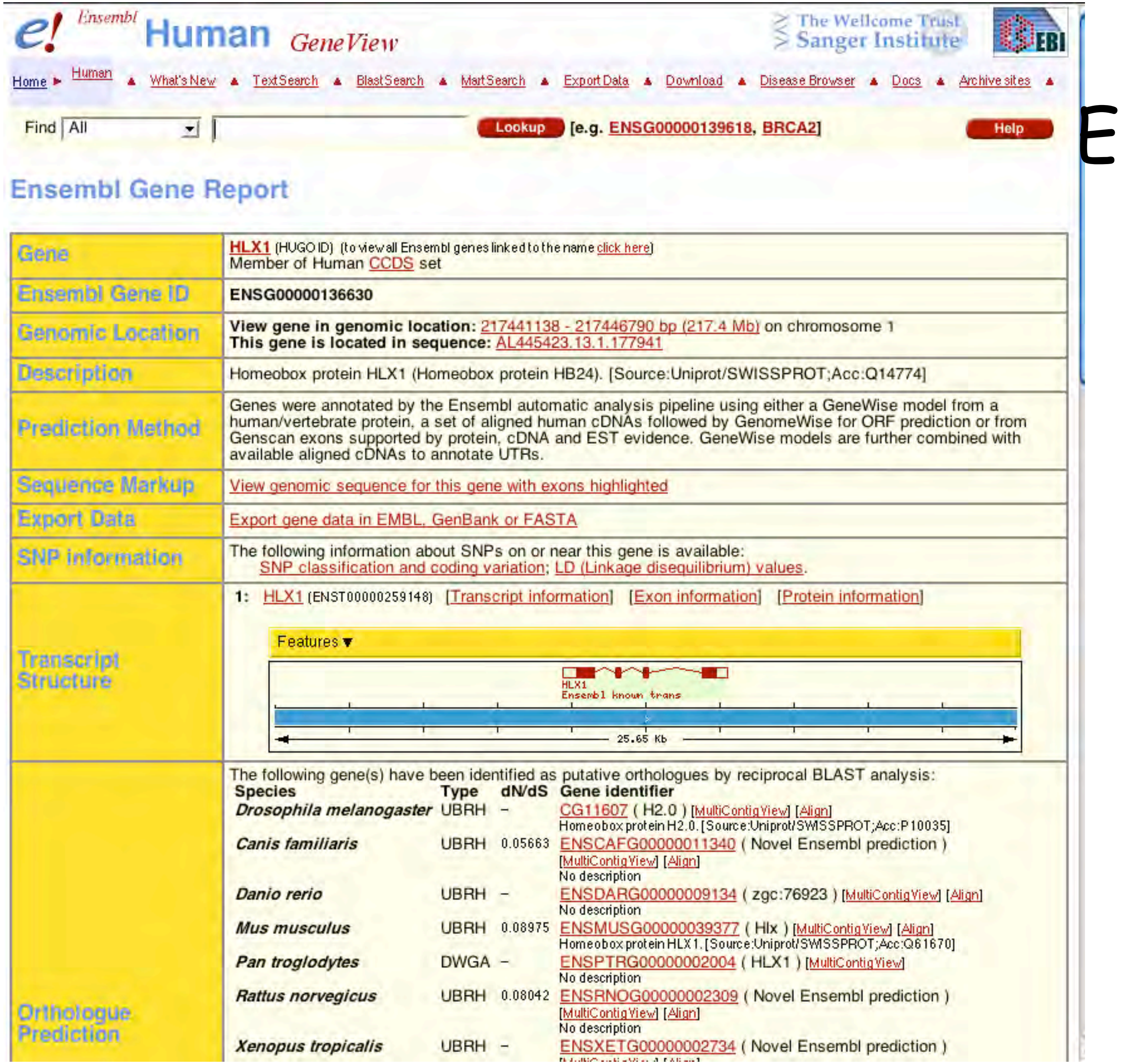
For context-sensitive help on any web page click

Questions or suggestions? Try

Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site)

This release is based on the NCBI 35 assembly of the human genome.
View the status history of the human assemblies.
Last Update: 14-12-2004
Ensembl gene predictions: 24194
(incl. 1978 pseudogenes)
Genscan gene predictions: 68101
Ensembl gene exons: 245215
Ensembl gene transcripts: 35838
Contigs: 27288
Clones: 27262
SuperContigs: 369
Chromosomes: 25 (plus 86 other fragments)
Base Pairs: 3272187692
Golden Path Length: 2863476365

EnsEMBL



[e.g. ENSG00000139618, BRCA2]

Ensembl Gene Report

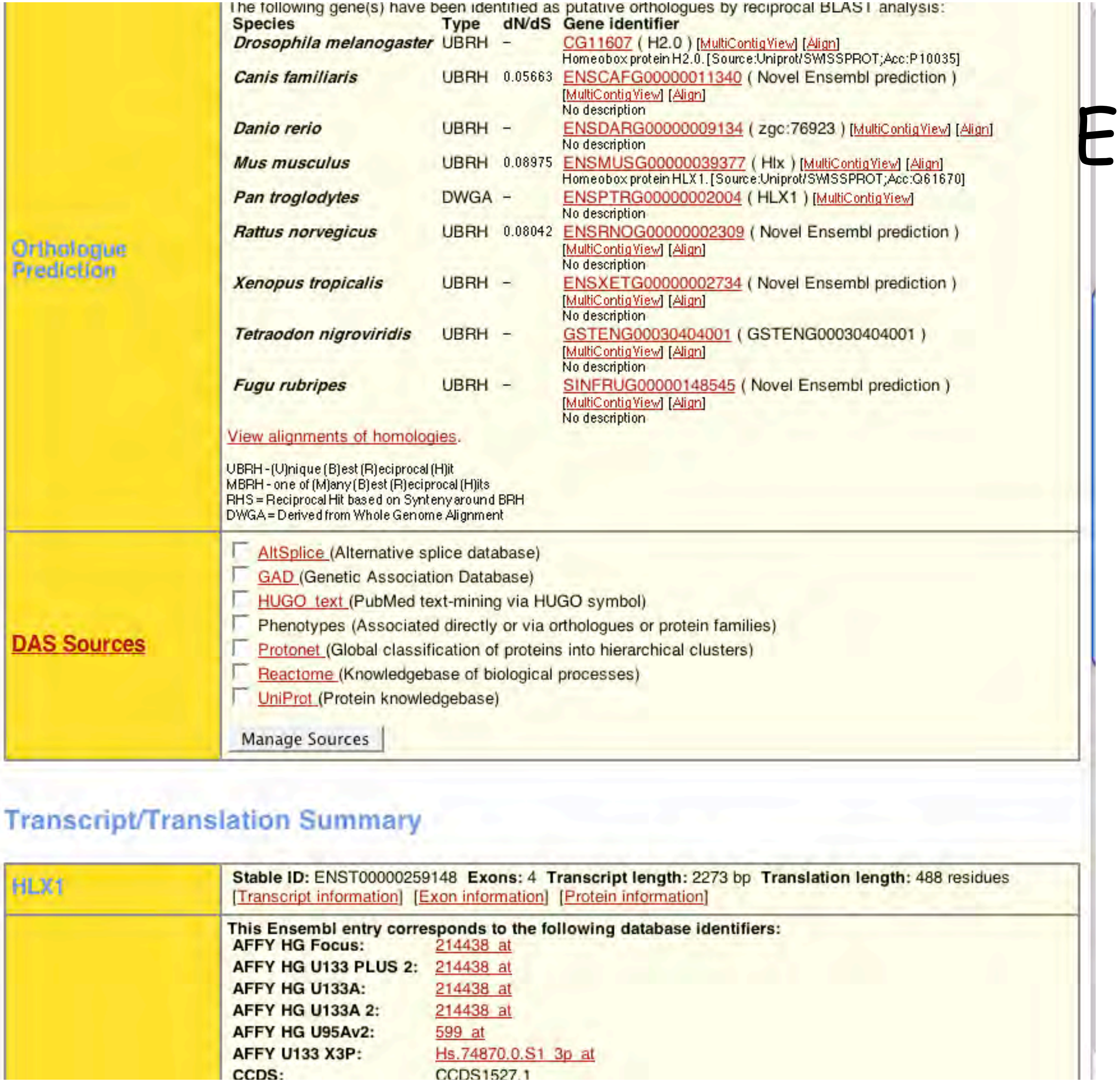
Gene	**HLX1** (HUGO ID) (to view all Ensembl genes linked to the name click here) Member of Human CCDS set
Ensembl Gene ID	**ENSG00000136630**
Genomic Location	**View gene in genomic location:** 217441138 - 217446790 bp (217.4 Mb) on chromosome 1 **This gene is located in sequence:** AL445423.13.1.177941
Description	Homeobox protein HLX1 (Homeobox protein HB24). [Source:Uniprot/SWISSPROT;Acc:Q14774]
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise model from a human/vertebrate protein, a set of aligned human cDNAs followed by GenomeWise for ORF prediction or from Genscan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.
Sequence Markup	View genomic sequence for this gene with exons highlighted
Export Data	Export gene data in EMBL, GenBank or FASTA
SNP information	The following information about SNPs on or near this gene is available: SNP classification and coding variation; LD (Linkage disequilibrium) values.
Transcript Structure	**1:** HLX1 [ENST00000259148] [Transcript information] [Exon information] [Protein information]
Orthologue Prediction	The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis:

Species	Type	dN/dS	Gene identifier
Drosophila melanogaster	UBRH	–	CG11607 (H2.0) [MultiContigView] [Align] Homeobox protein H2.0. [Source:Uniprot/SWISSPROT;Acc:P10035]
Canis familiaris	UBRH	0.05663	ENSCAFG00000011340 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Danio rerio	UBRH	–	ENSDARG00000009134 (zgc:76923) [MultiContigView] [Align] No description
Mus musculus	UBRH	0.08975	ENSMUSG00000039377 (Hlx) [MultiContigView] [Align] Homeobox protein HLX1. [Source:Uniprot/SWISSPROT;Acc:Q61670]
Pan troglodytes	DWGA	–	ENSPTRG00000002004 (HLX1) [MultiContigView] No description
Rattus norvegicus	UBRH	0.08042	ENSRNOG00000002309 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Xenopus tropicalis	UBRH	–	ENSXETG00000002734 (Novel Ensembl prediction)

1enh

EnsEMBL

1enh



Orthologue Prediction

The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis:

Species	Type	dN/dS	Gene identifier
Drosophila melanogaster	UBRH	–	CG11607 (H2.0) [MultiContigView] [Align] Homeobox protein H2.0. [Source:Uniprot/SWISSPROT;Acc:P10035]
Canis familiaris	UBRH	0.05663	ENSCAFG00000011340 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Danio rerio	UBRH	–	ENSDARG00000009134 (zgc:76923) [MultiContigView] [Align] No description
Mus musculus	UBRH	0.08975	ENSMUSG00000039377 (Hlx) [MultiContigView] [Align] Homeobox protein HLX1. [Source:Uniprot/SWISSPROT;Acc:Q61670]
Pan troglodytes	DWGA	–	ENSPTRG00000002004 (HLX1) [MultiContigView] No description
Rattus norvegicus	UBRH	0.08042	ENSRNOG00000002309 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Xenopus tropicalis	UBRH	–	ENSXETG00000002734 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Tetraodon nigroviridis	UBRH	–	GSTENG00030404001 (GSTENG00030404001) [MultiContigView] [Align] No description
Fugu rubripes	UBRH	–	SINFRUG00000148545 (Novel Ensembl prediction) [MultiContigView] [Align] No description

View alignments of homologies.

UBRH - (U)nique (B)est (R)eciprocal (H)it
MBRH - one of (M)any (B)est (R)eciprocal (H)its
RHS = Reciprocal Hit based on Synteny around BRH
DWGA = Derived from Whole Genome Alignment

DAS Sources

- ☐ AltSplice (Alternative splice database)
- ☐ GAD (Genetic Association Database)
- ☐ HUGO text (PubMed text-mining via HUGO symbol)
- ☐ Phenotypes (Associated directly or via orthologues or protein families)
- ☐ Protonet (Global classification of proteins into hierarchical clusters)
- ☐ Reactome (Knowledgebase of biological processes)
- ☐ UniProt (Protein knowledgebase)

Transcript/Translation Summary

HLX1

Stable ID: ENST00000259148 **Exons:** 4 **Transcript length:** 2273 bp **Translation length:** 488 residues
[Transcript information] [Exon information] [Protein information]

This Ensembl entry corresponds to the following database identifiers:

AFFY HG Focus:	214438 at
AFFY HG U133 PLUS 2:	214438 at
AFFY HG U133A:	214438 at
AFFY HG U133A 2:	214438 at
AFFY HG U95Av2:	599 at
AFFY U133 X3P:	Hs.74870.0.S1 3p at
CCDS:	CCDS1527.1

EnsEMBL

1enh

Similarity Matches

This Ensembl entry corresponds to the following database identifiers:

AFFY HG Focus:	214438_at
AFFY HG U133 PLUS 2:	214438_at
AFFY HG U133A:	214438_at
AFFY HG U133A 2:	214438_at
AFFY HG U95Av2:	599_at
AFFY U133 X3P:	Hs.74870.0.S1_3p_at
CCDS:	CCDS1527.1
EMBL:	AF217621 [align] AL445423 [align] BC007294 [align] BC033808 [align] M60721 [align] S56767 [align] U14325 [align] U14326 [align] U14327 [align] U14328 [align]
EntrezGene:	3142
HUGO:	Search GeneCards for HLX1
IPI:	IPI00030354.1 IPI00292164.2 IPI00550758.1
MIM:	142995
Protein ID:	AAC51346.1 [align] AAD13883.1 [align] AAF65541.1 [align] AAH07294.1 [align] AAH33808.1 [align] CAH72120.1 [align]
RefSeq dna:	NM_021958.2 [Target %id: 99; Query %id: 97] [align]
RefSeq peptide:	NP_068777.1 [Target %id: 100; Query %id: 100] [align]
Sequence Publications:	1672660 1676597 22388257 7806220 8094082 8095486 91291757
UniProt/Swiss-Prot:	Q14774 [Target %id: 99; Query %id: 99] [align]
UniProt/TrEMBL:	Q9NZ75 [Target %id: 100; Query %id: 100] [align]

GO

The following GO terms have been mapped to this entry via UniProt:

GO:0003700 [transcription factor activity] IEA
GO:0005634 [nucleus] IEA
GO:0006355 [regulation of transcription, DNA-dependent] IEA
GO:0006366 [transcription from Pol II promoter] NR
GO:0007275 [development] TAS

InterPro

IPR001356 Homeobox - [View other EnsEMBL genes with this domain]
IPR000047 Helix-turn-helix motif, lambda-like repressor - [View other EnsEMBL genes with this domain]

Protein Family

ENSF00000003396 : HOMEOBOX
This cluster contains 1 Ensembl gene member(s)

Transcript Structure

Protein Features

Banques spécialisées

CAZy
http://afmb.cnrs-mrs.fr/CAZY/index.html

HIV
http://www.hiv.lanl.gov/content/hiv-db/MAP/hivmap.html

A. Thaliana
http://www.hiv.lanl.gov/content/hiv-db/MAP/hivmap.html

CAZy - Carbohydrate-Active enZYmes

Last updated on 2005 Mar 07

The CAZy database describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds.

CAZymes per Class

Glycosidases and Transglycosidases

Glycosyltransferases

Polysaccharide Lyases

Carbohydrate Esterases

Carbohydrate-Binding Modules

CAZymes per Organism

Access by completely sequenced organism

Search

Banque A. thaliana

About TAIR | Sitemap | Contact | Help | Order | Login | Logout

tair The Arabidopsis Information Resource

TAIR Database ▾ Quick Search

Advanced Search
- Genes UPDATE
- Proteins
- Markers
- DNA
- Germplasm
- Ecotype
- Polymorphism/Allele
- People/Labs
- Protocols
- Publications
- Sequences
- Keyword Browser
- GO Annotations
- Locus History
- Microarray Elements
- Microarray Experiments
- Microarray Expression
- More....

Stocks
- About ABRC
- Place Order
- ABRC Catalog
- Search Seed Stocks
- Search DNA Stocks
- Search Order History
- Search My Stock Orders
- More....

News
- TAIR News
- Newsgroup
- Conferences..
- Job Postings

Analysis Tools
- SeqViewer
- MapViewer
- AraCyc Pathways
- BLAST
- WU-BLAST2
- FASTA
- Patmatch
- Microarray Analysis
- Motif Analysis
- Bulk Downloads UPDATE
- Chromosome Map Tool
- More....

External Links
- Stock Centers..
- Insertion, Knockout & Other Mutations
- Nomenclature
- Sequence Analysis
- Genome Databases
- Proteome Resources
- Cis-Element Resources
- Microarrays
- More....

Arabidopsis Info
- About Arabidopsis
- Genome Initiative
- Functional Genomics
- Monsanto SNPs & Ler
- Gene Expression
- Education & Outreach
- Gene Families
- Gene Class Symbols
- Ontologies
- Data Submission
- Protocols & LabManuals
- More....

FTP Downloads
- Sequences
- Genes
- Maps
- Microarrays
- Proteins
- Pathways
- Ontologies
- Protocols
- Data Submission NEW
- User Requests
- More....

Breaking News

ABRC/NASC Request Donations
[Mar 1, 2005]
ABRC and NASC are continuing the campaign to procure published mutants. Donations are requested! Details/submission info here.

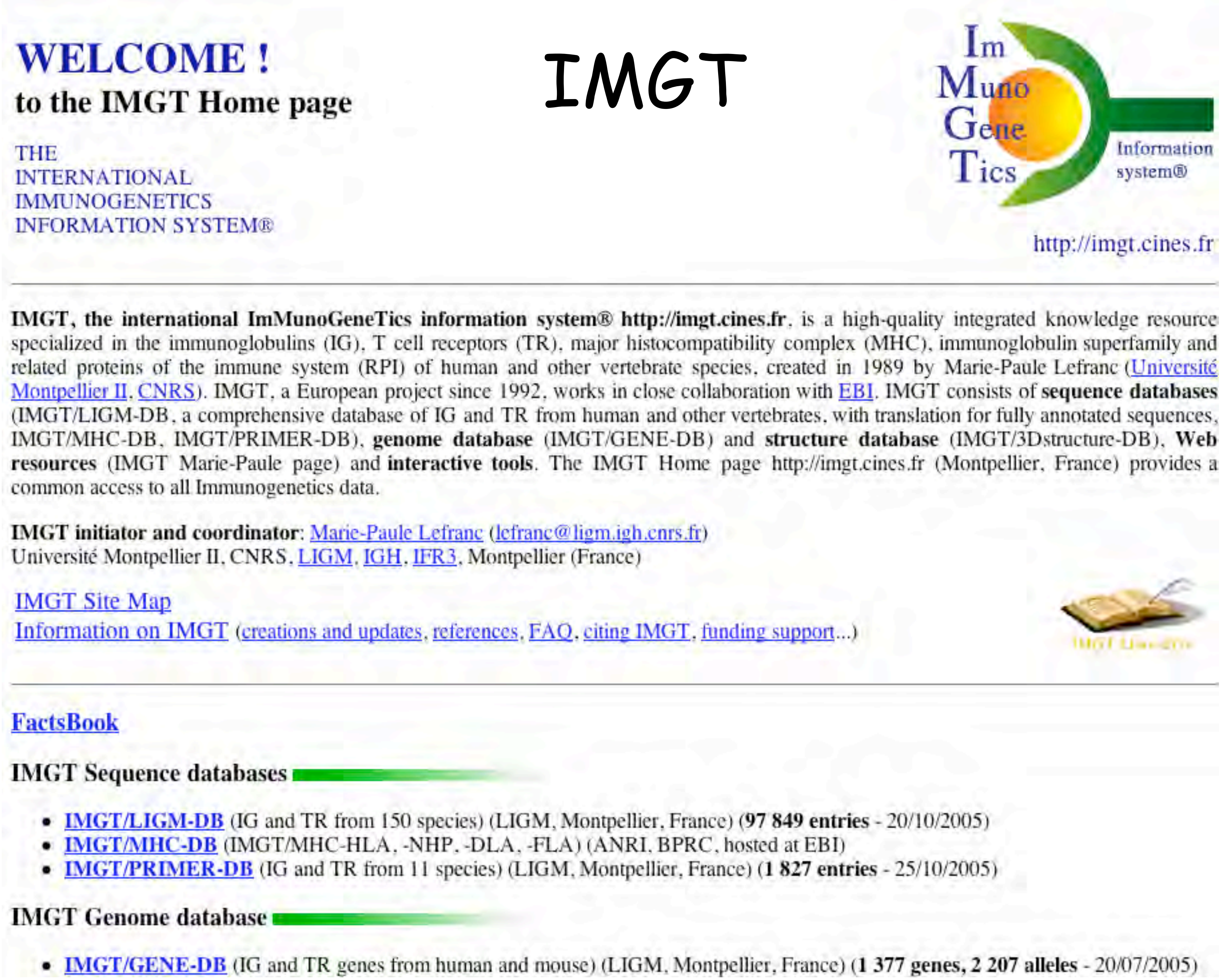
TAIR Metabolic Pathway Curator Position Available
[Feb 22, 2005]
Metabolic Pathway Curator position available immediately.

TAIR Dec 2004/Jan 2005 News Release
[Feb 8, 2005]
Get the most recent TAIR News.

TAIR Genome Annotation Curator Position Available
[Jan 25, 2005]
Genome Sequence Curator position available July 1, 2005.

16th International Conference on Arabidopsis Research, June 15-19, 2005

IMGT



WELCOME !

to the IMGT Home page

THE
INTERNATIONAL
IMMUNOGENETICS
INFORMATION SYSTEM®

http://imgt.cines.fr

IMGT, the international ImMunoGeneTics information system® http://imgt.cines.fr, is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC), immunoglobulin superfamily and related proteins of the immune system (RPI) of human and other vertebrate species, created in 1989 by Marie-Paule Lefranc (Université Montpellier II, CNRS). IMGT, a European project since 1992, works in close collaboration with EBI. IMGT consists of **sequence databases** (IMGT/LIGM-DB, a comprehensive database of IG and TR from human and other vertebrates, with translation for fully annotated sequences, IMGT/MHC-DB, IMGT/PRIMER-DB), **genome database** (IMGT/GENE-DB) and **structure database** (IMGT/3Dstructure-DB), **Web resources** (IMGT Marie-Paule page) and **interactive tools**. The IMGT Home page http://imgt.cines.fr (Montpellier, France) provides a common access to all Immunogenetics data.

IMGT initiator and coordinator: Marie-Paule Lefranc (lefranc@ligm.igh.cnrs.fr)
Université Montpellier II, CNRS, LIGM, IGH, IFR3, Montpellier (France)

IMGT Site Map
Information on IMGT (creations and updates, references, FAQ, citing IMGT, funding support...)

FactsBook

IMGT Sequence databases

- **IMGT/LIGM-DB** (IG and TR from 150 species) (LIGM, Montpellier, France) (**97 849 entries** - 20/10/2005)
- **IMGT/MHC-DB** (IMGT/MHC-HLA, -NHP, -DLA, -FLA) (ANRI, BPRC, hosted at EBI)
- **IMGT/PRIMER-DB** (IG and TR from 11 species) (LIGM, Montpellier, France) (**1 827 entries** - 25/10/2005)

IMGT Genome database

- **IMGT/GENE-DB** (IG and TR genes from human and mouse) (LIGM, Montpellier, France) (**1 377 genes, 2 207 alleles** - 20/07/2005)

Recherche d'information dans les banques

- **Lexicographique** : recherche de fragments de texte strictement présents dans la banque

- **Avec valeur ajoutée** : recherche des informations voisines, des synonymes. Criblage de banque pour rechercher toute la famille d'une séquence donnée

Interrogation littérale

Interrogation sur plusieurs bases à la fois; possibilité utiliser les références croisées (rechercher dans EMBL les séquences issues d'une interrogation de SW)

Opérateurs logiques : ET, OU et NON

Logiciel SRS installé sur de nombreux serveurs Web (SW, NCBI, EBI, PFAM…). Interface graphique commune

Liste des serveurs disponible à http://downloads.lionbio.co.uk/publicsrs.html

SRS

Sequence Retrieval System

Network Browser for Databanks in Molecular Biology

Start a new SRS session

The SRS Manual

List of all SRS5 servers

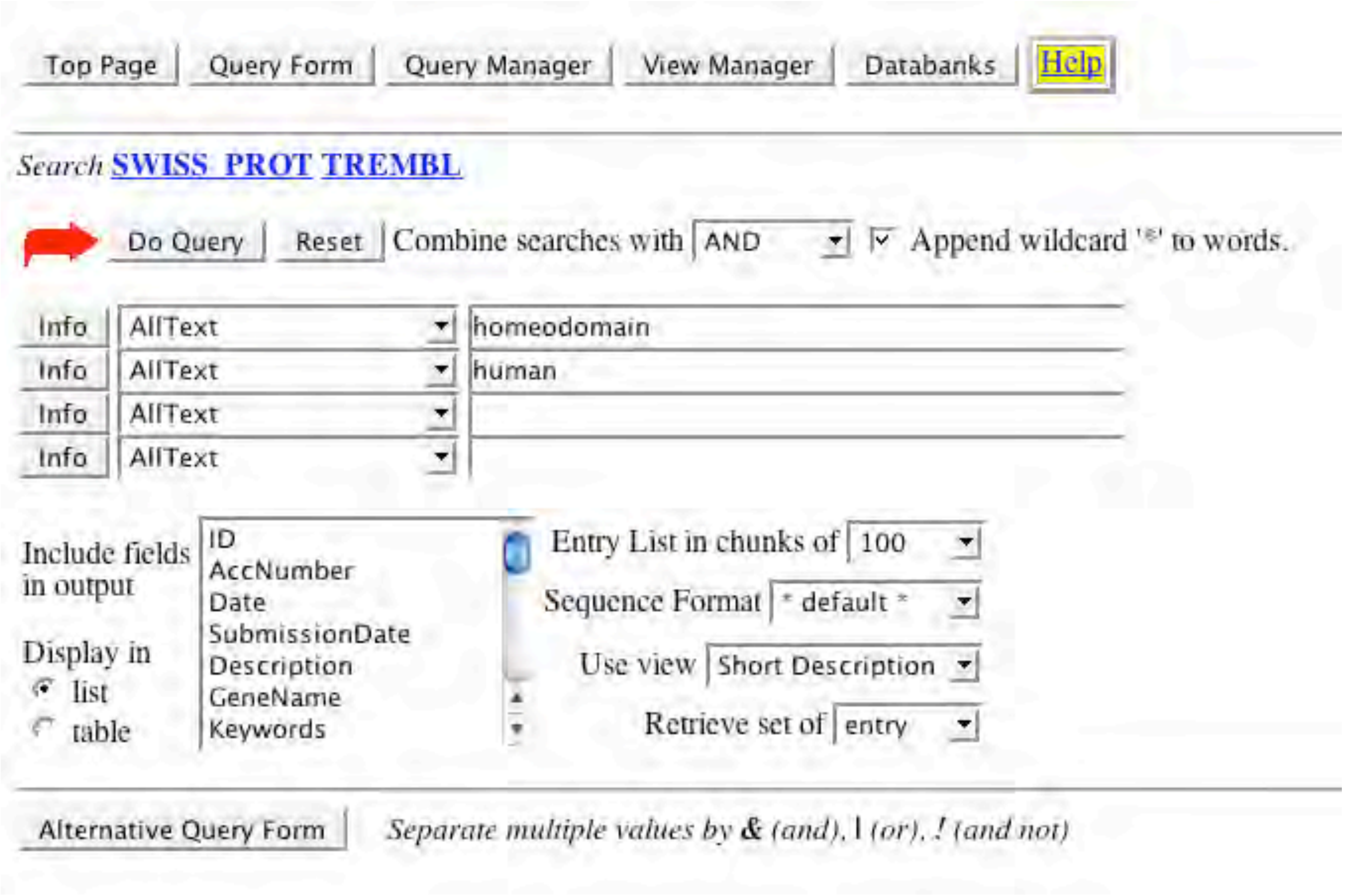
The SRS newsgroup

Back to the ExPASy home page

This is the SRSWWW server at ExPASy, version 5.1.0

Hosted by APAF Australia Mirror sites: Bolivia Brazil Canada China Korea Switzerland Taiwan USA

SRS



Search SWISS_PROT TREMBL

Combine searches with AND ☑ Append wildcard '*' to words.

AllText homeodomain
AllText human
AllText
AllText

Include fields in output
ID
AccNumber
Date
SubmissionDate
Description
GeneName
Keywords

Display in
● list
○ table

Entry List in chunks of 100
Sequence Format * default *
Use view Short Description
Retrieve set of entry

Separate multiple values by & (and), | (or), ! (and not)

SRS3D

Query forms

▸ bookmark this link to return to your session

If you find problems or have suggestions please mail the SRS administrator

show all ⊞ collapse all ⊟

⊟ **Sequence libraries - complete**

all ☐ UniProt

⊟ **Sequence libraries - subsections**

all ☐ UniProt/Swiss-Prot ☐ UniProt/TrEMBL ☐ UniProt/TrEMBLNew

⊟ **InterPro&Related**

all ☐ INTERPRO ☐ IPRMATCHES

⊟ **SeqRelated**

all ☐ CustomAnnot

⊟ **User Owned Databanks**

all ☐ USERPROTEIN

⊟ **Application Results**

all ☐ BLASTP

⊟ **Macromolecular Structures**

all ☐ PDB ☐ DSSP ☐ HSSP ☐ FSSP

☐ FSSPmatches ☐ PSSH ☐ HSSPchain ☐ HSSPalign

⊟ **Others**

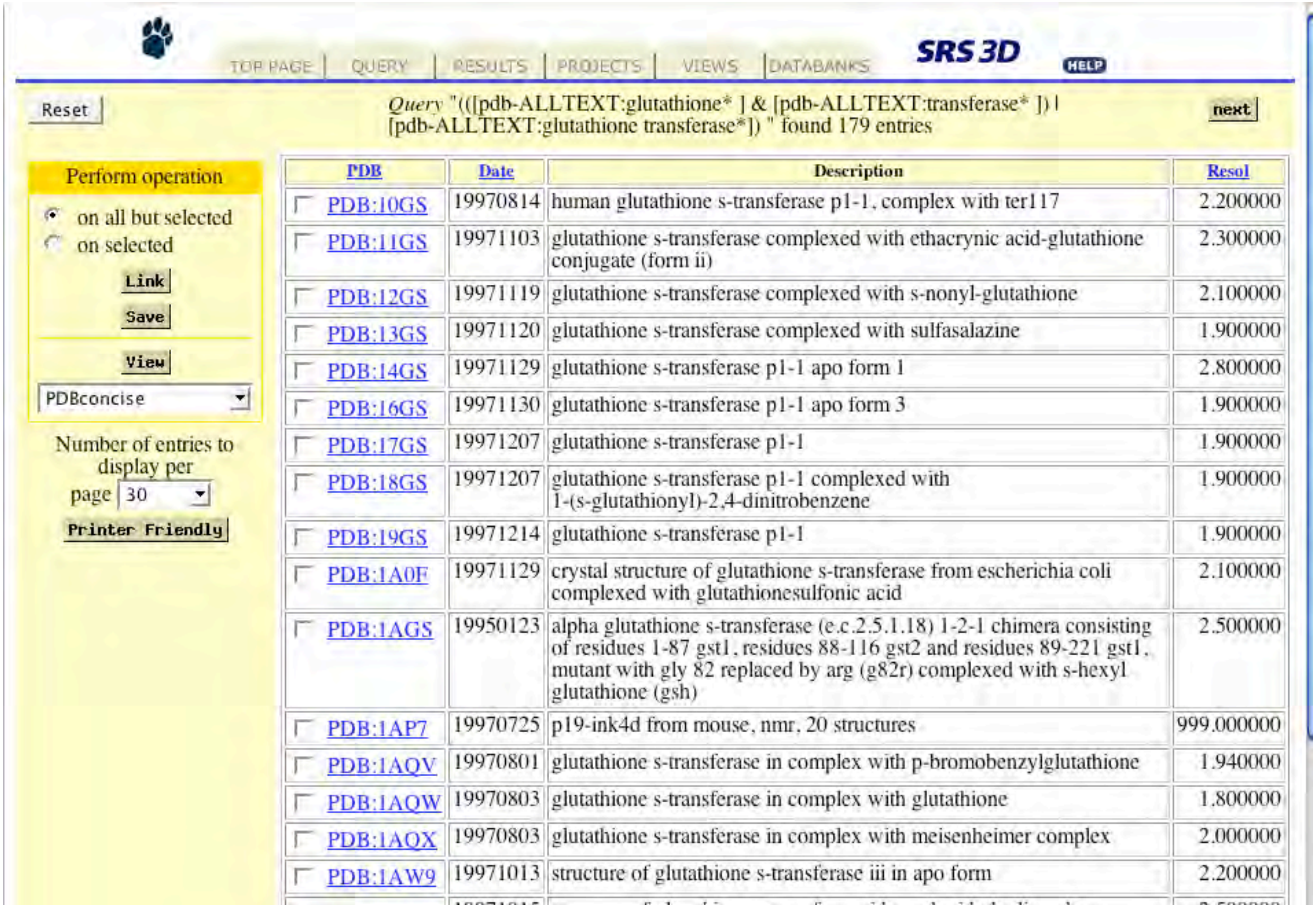
all ☐ PDBequiv

⊟ **System**

all ☐ PRISMASTATUS ☐ PRISMATIME

SRS 6.1.3.11 | feedback

SRS3D



Query "(([pdb-ALLTEXT:glutathione*] & [pdb-ALLTEXT:transferase*]) | [pdb-ALLTEXT:glutathione transferase*]) " found 179 entries

Perform operation

● on all but selected
○ on selected

PDBconcise

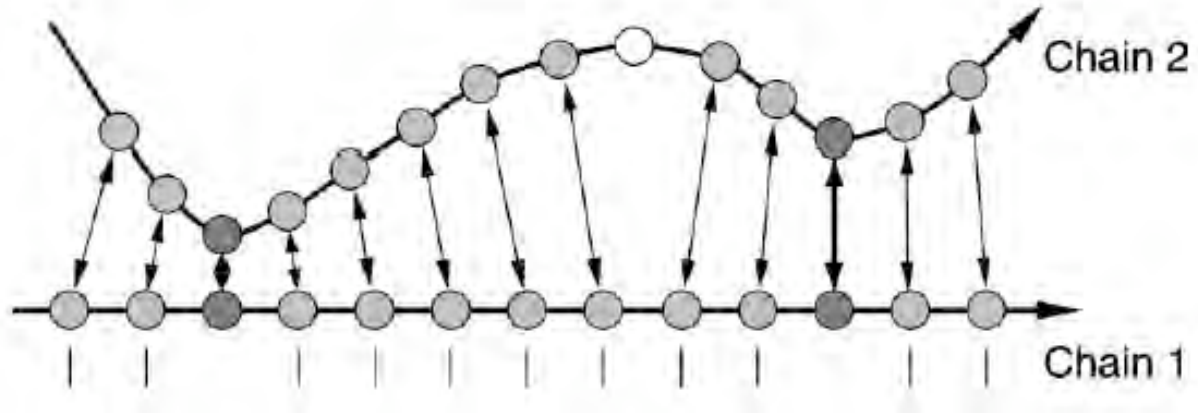
Number of entries to display per page 30

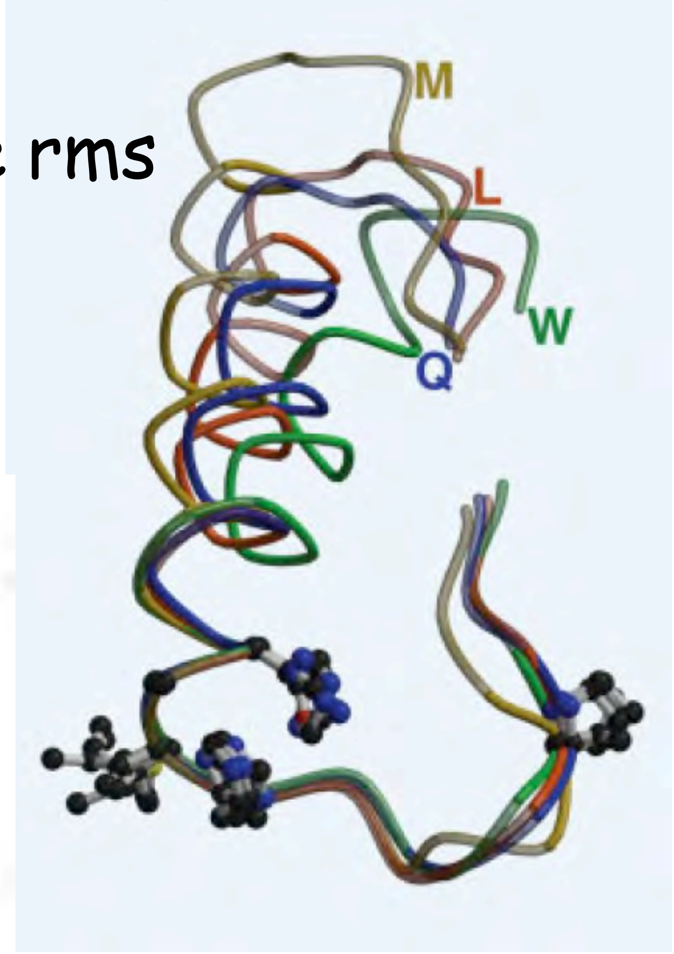
PDB	Date	Description	Resol
☐ PDB:10GS	19970814	human glutathione s-transferase p1-1, complex with ter117	2.200000
☐ PDB:11GS	19971103	glutathione s-transferase complexed with ethacrynic acid-glutathione conjugate (form ii)	2.300000
☐ PDB:12GS	19971119	glutathione s-transferase complexed with s-nonyl-glutathione	2.100000
☐ PDB:13GS	19971120	glutathione s-transferase complexed with sulfasalazine	1.900000
☐ PDB:14GS	19971129	glutathione s-transferase p1-1 apo form 1	2.800000
☐ PDB:16GS	19971130	glutathione s-transferase p1-1 apo form 3	1.900000
☐ PDB:17GS	19971207	glutathione s-transferase p1-1	1.900000
☐ PDB:18GS	19971207	glutathione s-transferase p1-1 complexed with 1-(s-glutathionyl)-2,4-dinitrobenzene	1.900000
☐ PDB:19GS	19971214	glutathione s-transferase p1-1	1.900000
☐ PDB:1A0F	19971129	crystal structure of glutathione s-transferase from escherichia coli complexed with glutathionesulfonic acid	2.100000
☐ PDB:1AGS	19950123	alpha glutathione s-transferase (e.c.2.5.1.18) 1-2-1 chimera consisting of residues 1-87 gst1, residues 88-116 gst2 and residues 89-221 gst1, mutant with gly 82 replaced by arg (g82r) complexed with s-hexyl glutathione (gsh)	2.500000
☐ PDB:1AP7	19970725	p19-ink4d from mouse, nmr, 20 structures	999.000000
☐ PDB:1AQV	19970801	glutathione s-transferase in complex with p-bromobenzylglutathione	1.940000
☐ PDB:1AQW	19970803	glutathione s-transferase in complex with glutathione	1.800000
☐ PDB:1AQX	19970803	glutathione s-transferase in complex with meisenheimer complex	2.000000
☐ PDB:1AW9	19971013	structure of glutathione s-transferase iii in apo form	2.200000

Comparaison des structures

- Pour comparer des structures il faut d'abord les superposer résidu par résidu selon la correspondance établie lors de l'alignement des séquences.
- Ecart quadratique moyen est le rms

$$rms(i,j) = \sqrt{\frac{\sum_{k=1}^{N}\left(\vec{r}_{i+k} - \vec{r}_{j+k}\right)^2}{N}}$$



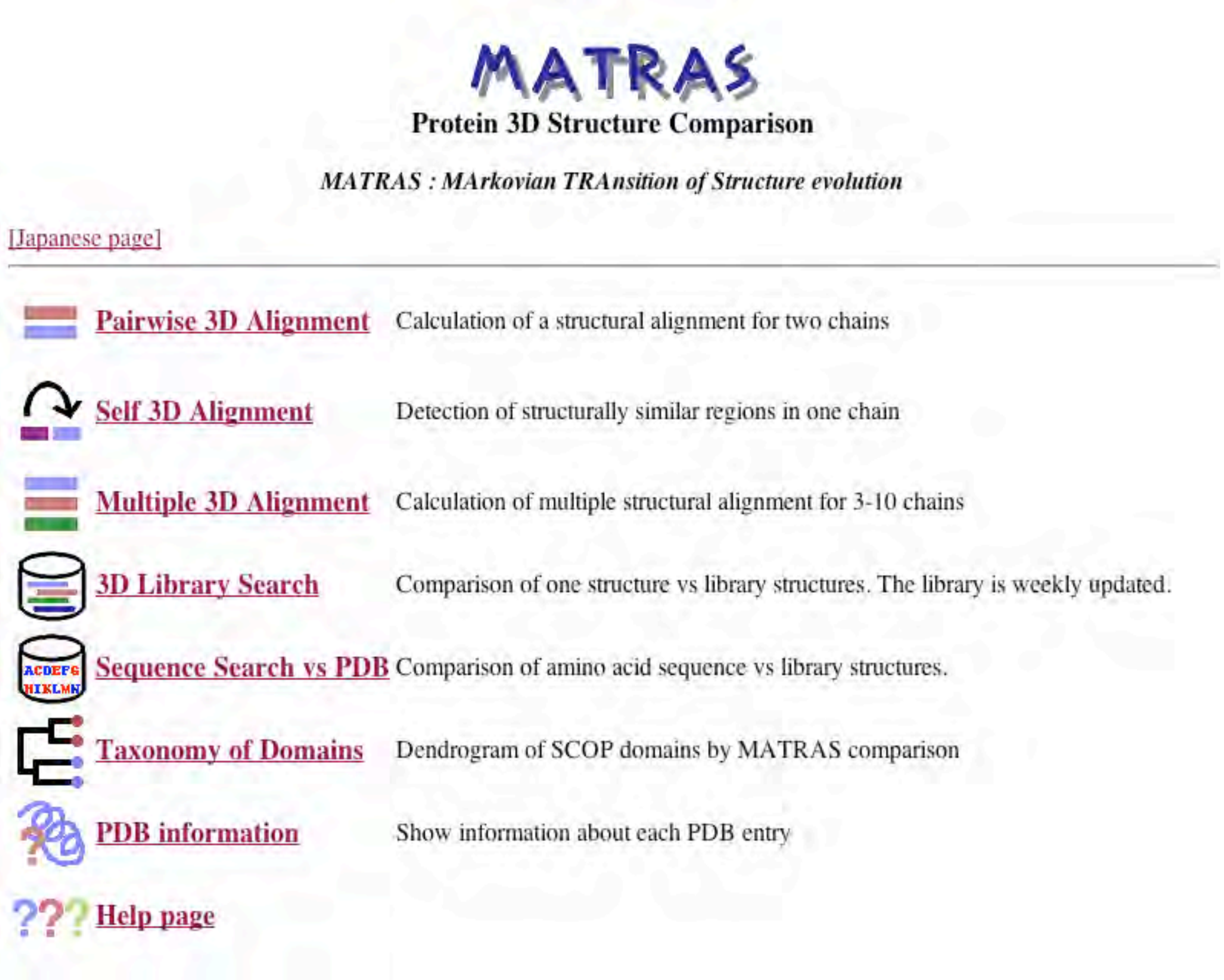


Superposer = comparer du 3D

- Nombreux sites

- MSD permet de sélectionner une partie de la structure à superposer
- http://www.ebi.ac.uk/msd-srv/ssm/cgi-bin/ssmserver
- MATRAS permet la même chose, et fournit une image
- http://biunit.aist-nara.ac.jp/matras/matras_pair.html
- Yakusa
- http://bioserv.rpbs.jussieu.fr/Yakusa/index.html

- Peu d'outils pour superposer un grand nombre de structures (Plus de 10): MUSTANG (pas en ligne)



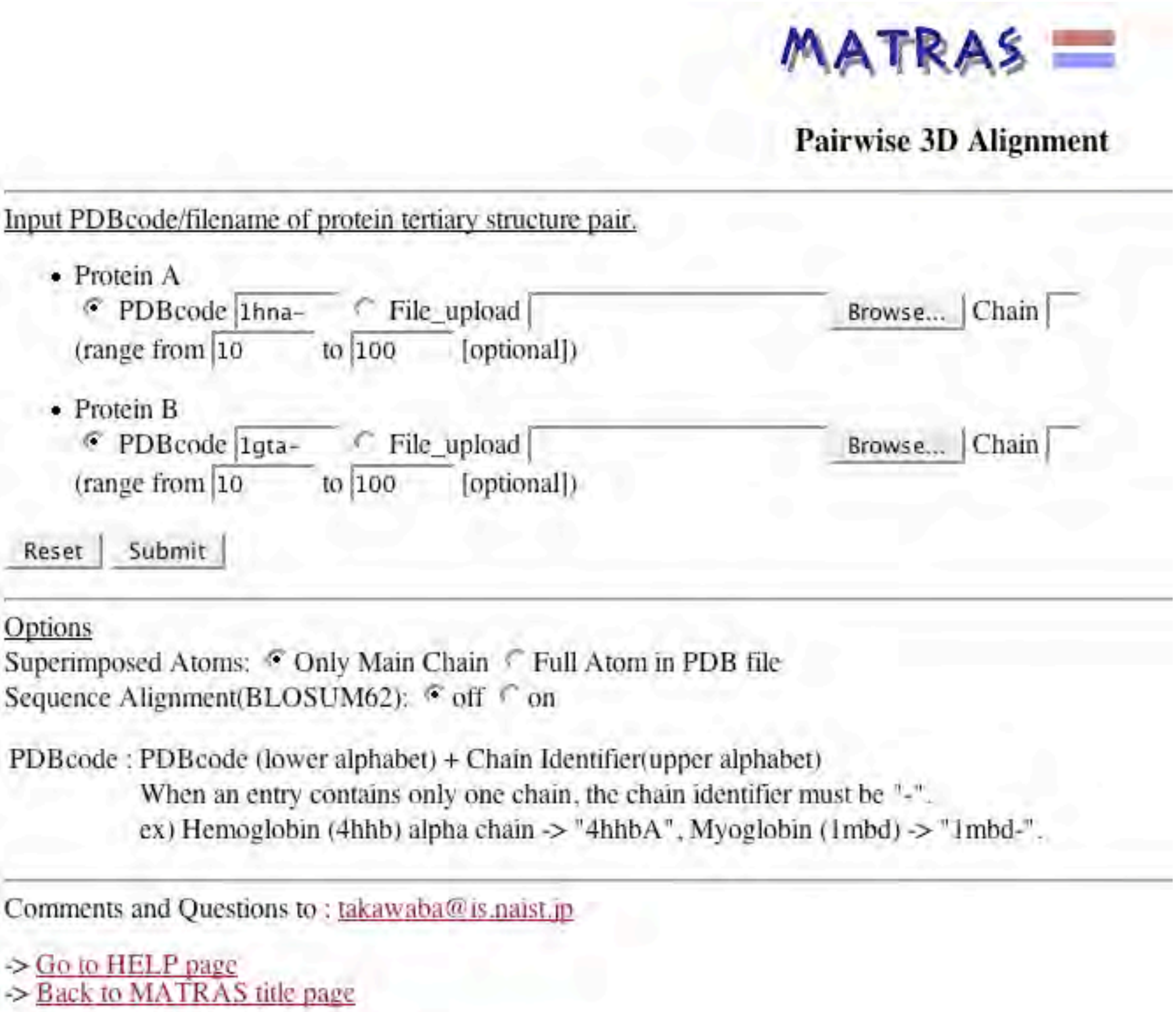
MATRAS

Protein 3D Structure Comparison

MATRAS : MArkovian TRAnsition of Structure evolution

[Japanese page]

Pairwise 3D Alignment	Calculation of a structural alignment for two chains
Self 3D Alignment	Detection of structurally similar regions in one chain
Multiple 3D Alignment	Calculation of multiple structural alignment for 3-10 chains
3D Library Search	Comparison of one structure vs library structures. The library is weekly updated.
Sequence Search vs PDB	Comparison of amino acid sequence vs library structures.
Taxonomy of Domains	Dendrogram of SCOP domains by MATRAS comparison
PDB information	Show information about each PDB entry
Help page	



Pairwise 3D Alignment

Input PDBcode/filename of protein tertiary structure pair.

- Protein A
  ● PDBcode 1hna- ○ File_upload Chain
  (range from 10 to 100 [optional])
- Protein B
  ● PDBcode 1gta- ○ File_upload Chain
  (range from 10 to 100 [optional])

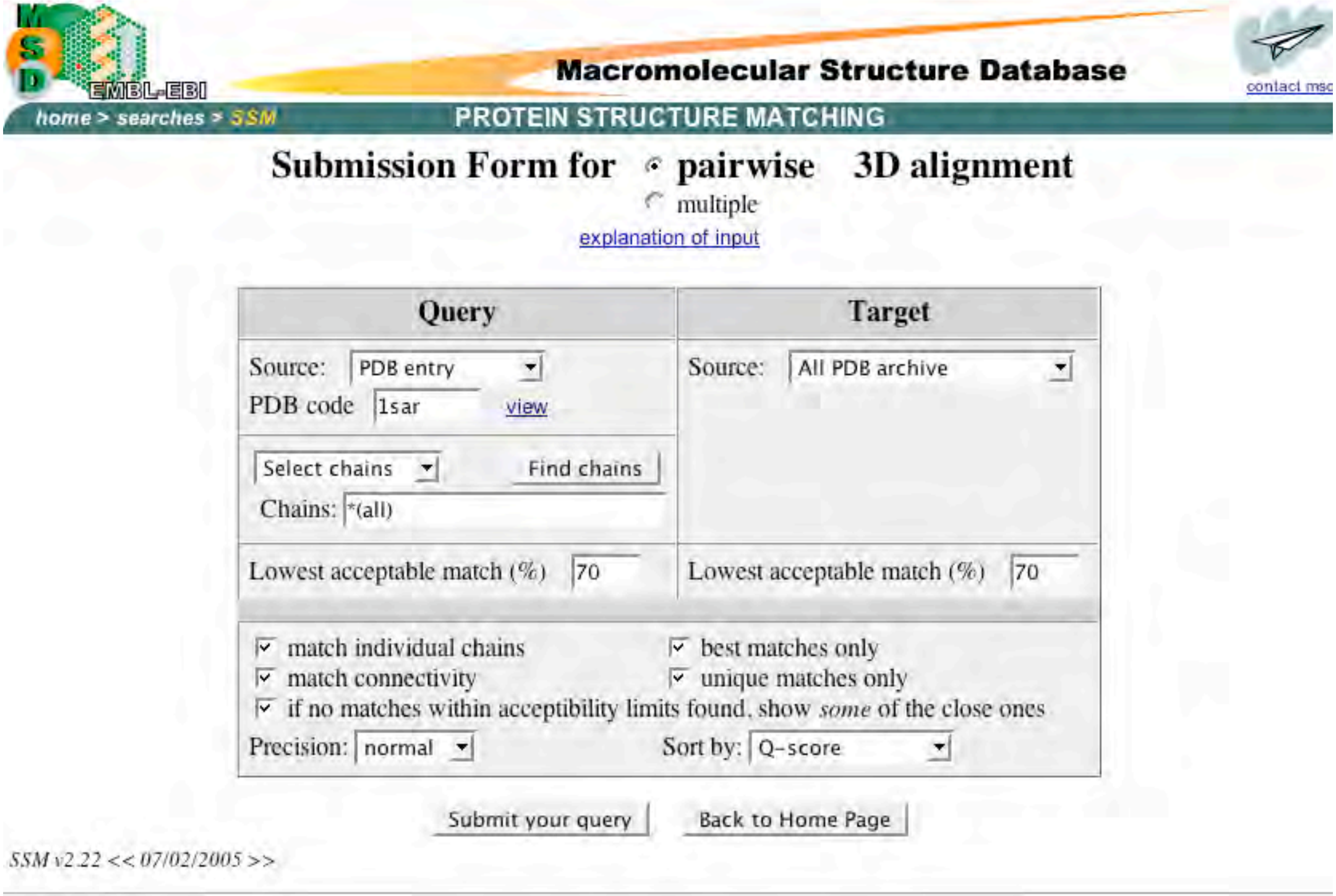
Options

Superimposed Atoms: ● Only Main Chain ○ Full Atom in PDB file
Sequence Alignment(BLOSUM62): ● off ○ on

PDBcode : PDBcode (lower alphabet) + Chain Identifier(upper alphabet)
When an entry contains only one chain, the chain identifier must be "-".
ex) Hemoglobin (4hhb) alpha chain -> "4hhbA", Myoglobin (1mbd) -> "1mbd-".

Comments and Questions to : takawaba@is.naist.jp

-> Go to HELP page
-> Back to MATRAS title page



Submission Form for ● pairwise ○ multiple 3D alignment

Query	Target
Source: PDB entry PDB code 1sar	Source: All PDB archive
Chains: *(all)	
Lowest acceptable match (%) 70	Lowest acceptable match (%) 70

☑ match individual chains
☑ match connectivity
☑ best matches only
☑ unique matches only
☑ if no matches within acceptibility limits found, show *some* of the close ones

Precision: normal

Sort by: Q-score

SSM v2.22 << 07/02/2005 >>

EBI-MSD

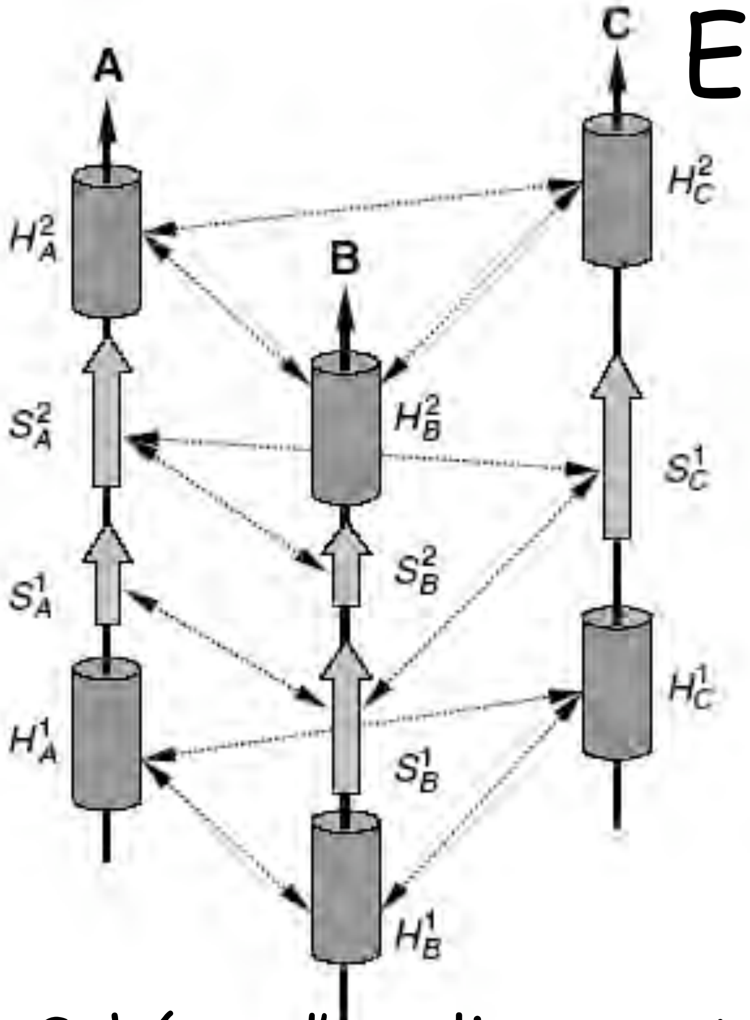
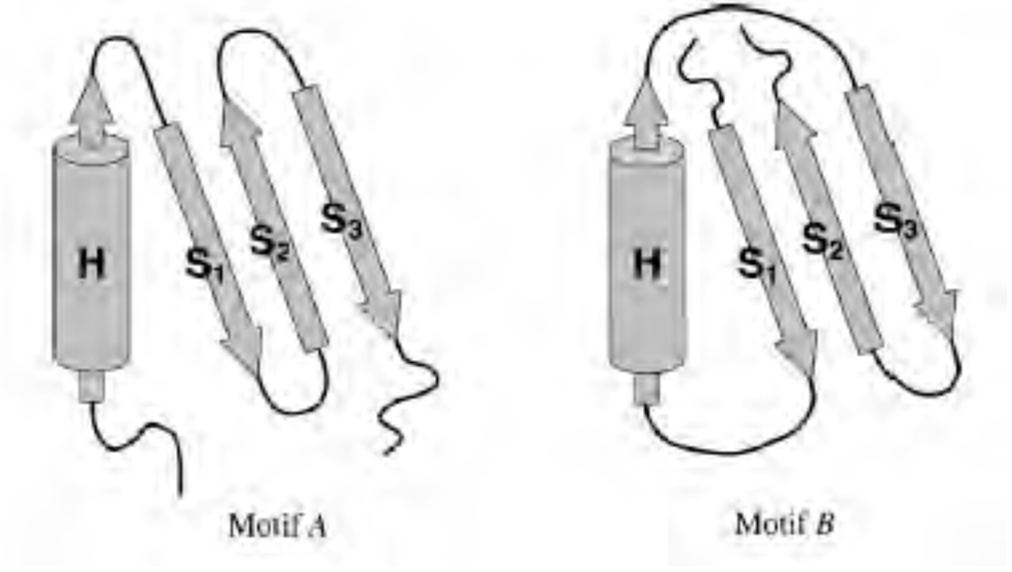


Schéma d'un alignement structural multiple. Les flèches dénotent les SSE alignés par paires.



Motifs A & B identiques en géométrie, mais connectiviés différentes. MSD gère connectivité entre les structures secondaires.

Krissinel (2004) Acta Cryst D60:2256

EBI-MSD

Fonction Q de qualité d'accord entre deux structures

rmsd entre les deux structures,

nombre de résidus alignés (Nalign) entre les deux séquences de longueur N1 et N2

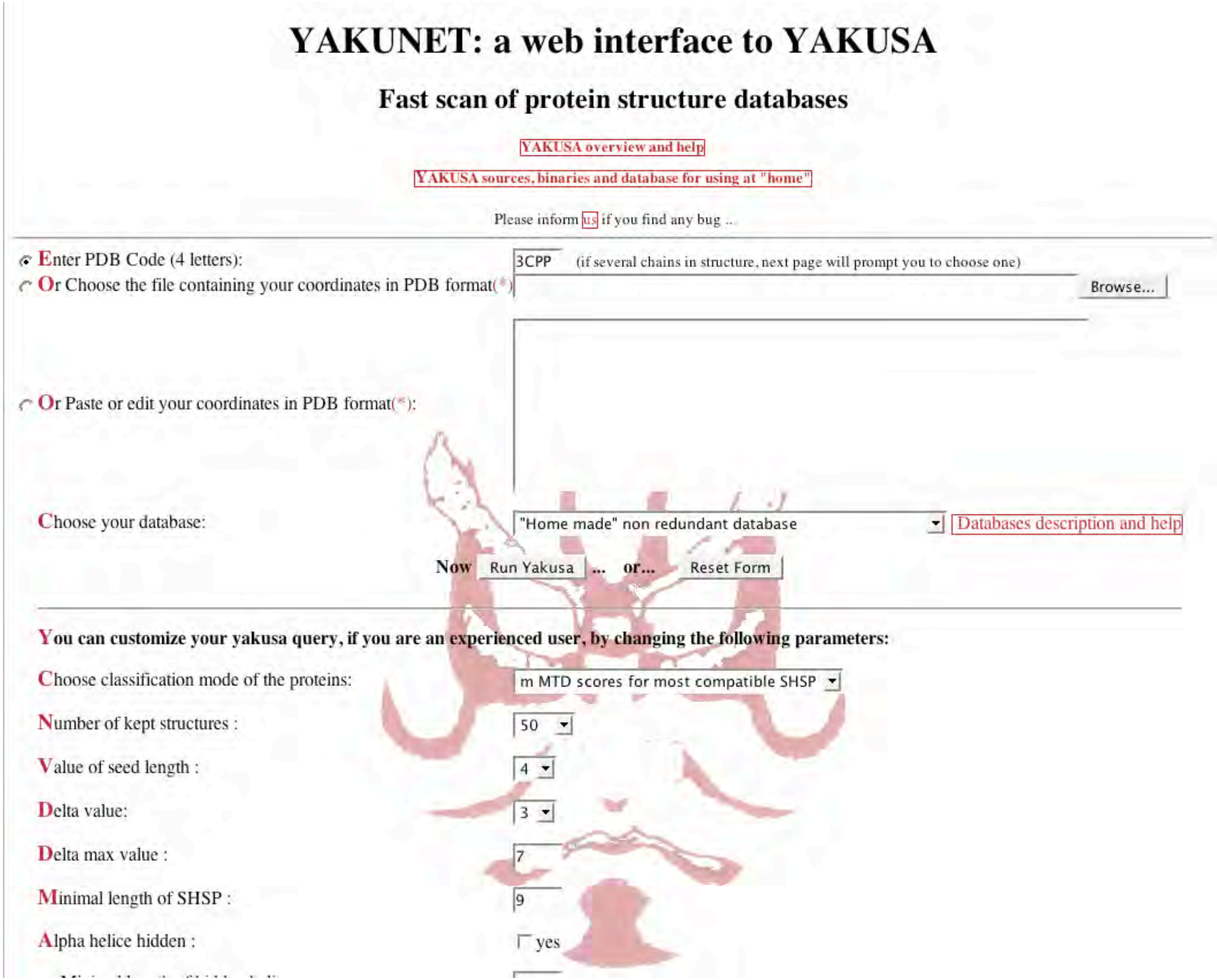
R° = paramètre empirique choisi à la valeur de 3 Å

Q = 1 pour structures identiques (N1=N2=Nalign) ; décroit vers zéro quand divergence augmente

Compromis entre rms faible et maximum de résidus alignés

$$Q_x^i = \sum_y \sum_j \frac{\left(N_{xy}^{ij}\right)^2}{\left(1 + \left(RMSD_{xy}^{ij}/R_0\right)^2\right) N_x^i N_y^j} \qquad Q = N_{\mathrm{align}}^2 / \{[1 + (\mathrm{RMSD}/R_0)^2] N_1 N_2\}$$

Krissinel (2004) Acta Cryst D60:2256



YAKUNET: a web interface to YAKUSA

Fast scan of protein structure databases

YAKUSA overview and help

YAKUSA sources, binaries and database for using at "home"

Please inform us if you find any bug ..

● Enter PDB Code (4 letters): 3CPP (if several chains in structure, next page will prompt you to choose one)

○ Or Choose the file containing your coordinates in PDB format(*)

○ Or Paste or edit your coordinates in PDB format(*):

Choose your database: "Home made" non redundant database

You can customize your yakusa query, if you are an experienced user, by changing the following parameters:

Choose classification mode of the proteins: m MTD scores for most compatible SHSP

Number of kept structures : 50

Value of seed length : 4

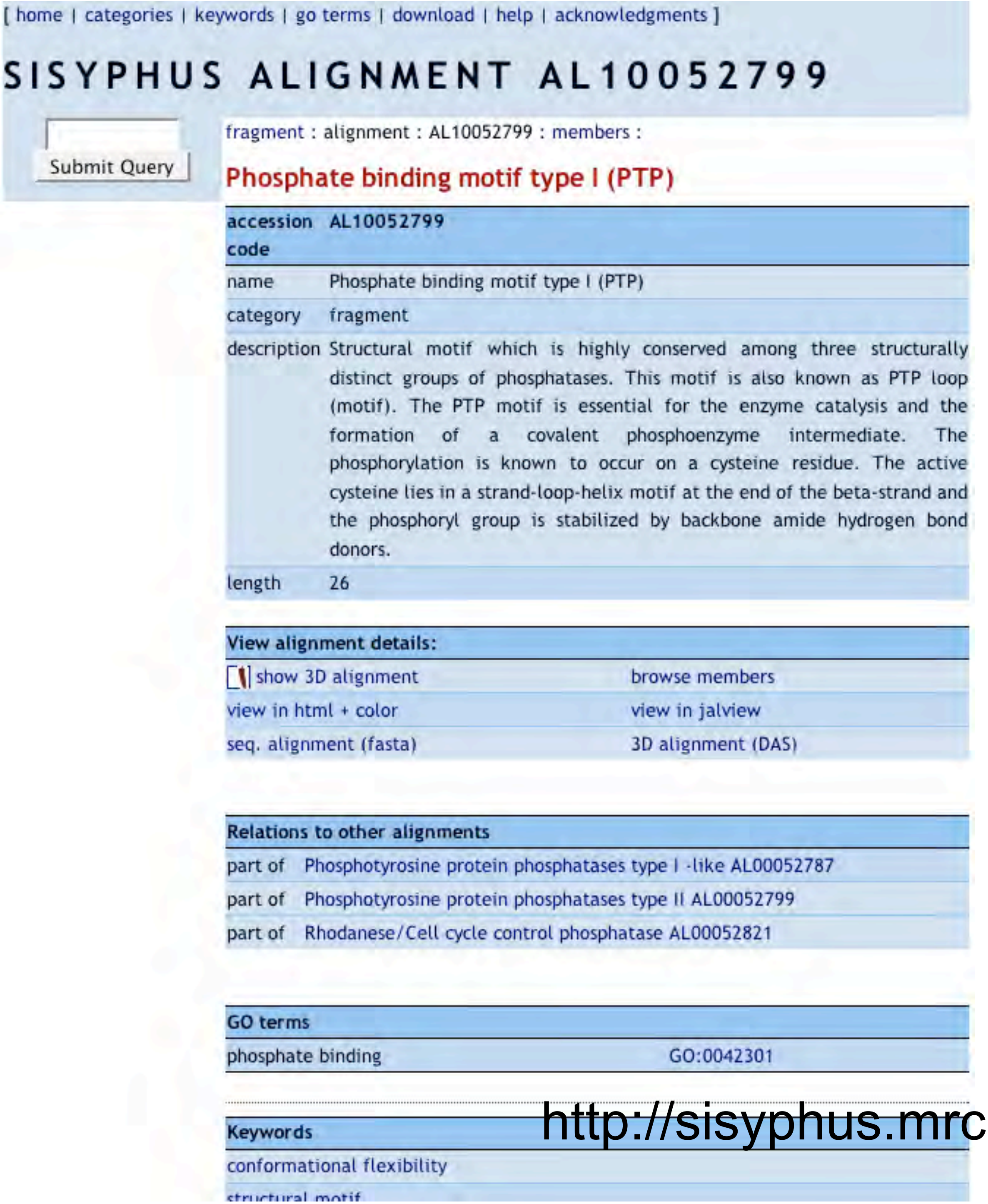
Delta value: 3

Delta max value : 7

Minimal length of SHSP : 9

Alpha helice hidden : ☐ yes

SISYPHUS



SISYPHUS ALIGNMENT AL10052799

fragment : alignment : AL10052799 : members :

Phosphate binding motif type I (PTP)

accession code	**AL10052799**
name	Phosphate binding motif type I (PTP)
category	fragment
description	Structural motif which is highly conserved among three structurally distinct groups of phosphatases. This motif is also known as PTP loop (motif). The PTP motif is essential for the enzyme catalysis and the formation of a covalent phosphoenzyme intermediate. The phosphorylation is known to occur on a cysteine residue. The active cysteine lies in a strand-loop-helix motif at the end of the beta-strand and the phosphoryl group is stabilized by backbone amide hydrogen bond donors.
length	26

View alignment details:	
show 3D alignment	browse members
view in html + color	view in jalview
seq. alignment (fasta)	3D alignment (DAS)

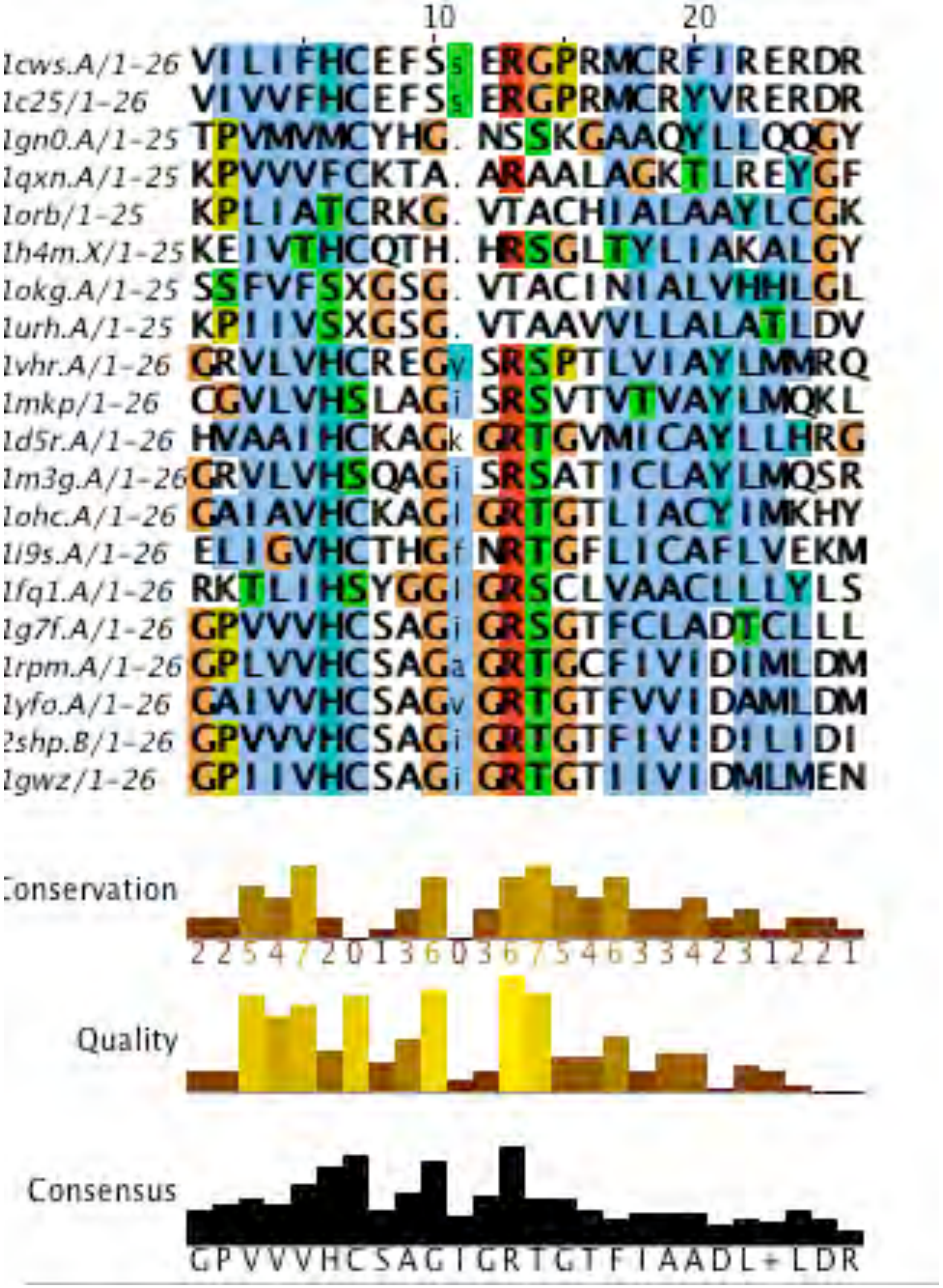
Relations to other alignments	
part of	Phosphotyrosine protein phosphatases type I -like AL00052787
part of	Phosphotyrosine protein phosphatases type II AL00052799
part of	Rhodanese/Cell cycle control phosphatase AL00052821

GO terms	
phosphate binding	GO:0042301

Keywords
conformational flexibility
structural motif

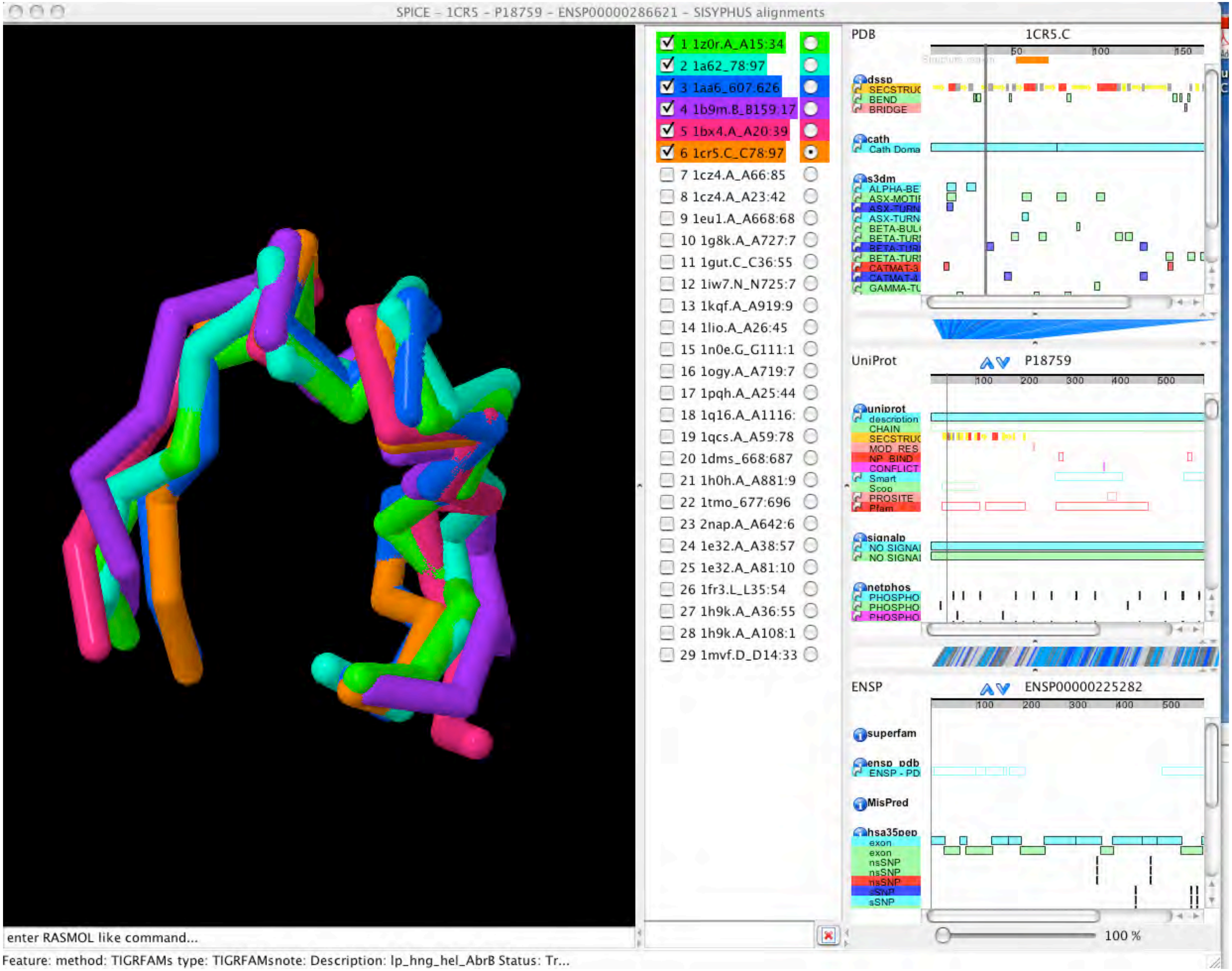
http://sisyphus.mrc-cpe.cam.ac.uk/sisyphus/

SISYPHUS



```
1cws.A/1-26 VILIFHCEFSsERGPRMCRFIRERDR
1c25/1-26   VIVVFHCEFSsERGPRMCRYVRERDR
1gn0.A/1-25 TPVMVMCYHG.NSSKGAAQYLLQQGY
1qxn.A/1-25 KPVVVFCKTA.ARAALAGKTLREYGF
1orb/1-25   KPLIATCRKG.VTACHIALAAYLCGK
1h4m.X/1-25 KEIVTHCQTH.HRSGLTYLIAKALGY
1okg.A/1-25 SSFVFSXGSG.VTACINIALVHHLGL
1urh.A/1-25 KPIIVSXGSG.VTAAVVLLALATLDV
1vhr.A/1-26 GRVLVHCREGySRSPTLVIAYLMMRQ
1mkp/1-26   CGVLVHSLAGiSRSVTVTVAYLMQKL
1d5r.A/1-26 HVAAIHCKAGkGRTGVMICAYLLHRG
1m3g.A/1-26 GRVLVHSQAGiSRSATICLAYLMQSR
1ohc.A/1-26 GAIAVHCKAGiGRTGTLIACYIMKHY
1i9s.A/1-26 ELIGVHCTHGfNRTGFLICAFLVEKM
1fq1.A/1-26 RKTLIHSYGGiGRSCLVAACLLLYLS
1g7f.A/1-26 GPVVVHCSAGiGRSGTFCLADTCLLL
1rpm.A/1-26 GPLVVHCSAGaGRTGCFIVIDIMLDM
1yfo.A/1-26 GAIVVHCSAGvGRTGTFVVIDAMLDM
2shp.B/1-26 GPVVVHCSAGiGRTGTFIVIDILIDI
1gwz/1-26   GPIIVHCSAGiGRTGTIIVIDMLMEN

Conservation 22547201360367546334231221
Consensus    GPVVVHCSAGIGRTGTFIAADL+LDR
```



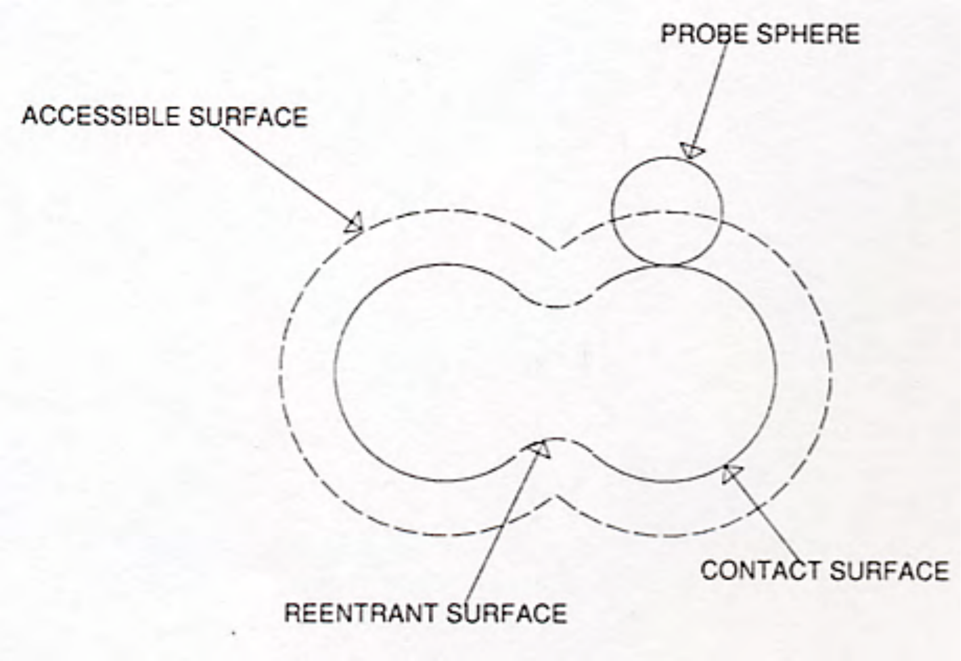
SPICE - 1CR5 - P18759 - ENSP00000286621 - SISYPHUS alignments

☑ 1 1z0r.A_A15:34 ○
☑ 2 1a62_78:97 ○
☑ 3 1aa6_607:626 ○
☑ 4 1b9m.B_B159:17 ○
☑ 5 1bx4.A_A20:39 ○
☑ 6 1cr5.C_C78:97 ●
☐ 7 1cz4.A_A66:85 ○
☐ 8 1cz4.A_A23:42 ○
☐ 9 1eu1.A_A668:68 ○
☐ 10 1g8k.A_A727:7 ○
☐ 11 1gut.C_C36:55 ○
☐ 12 1iw7.N_N725:7 ○
☐ 13 1kqf.A_A919:9 ○
☐ 14 1lio.A_A26:45 ○
☐ 15 1n0e.G_G111:1 ○
☐ 16 1ogy.A_A719:7 ○
☐ 17 1pqh.A_A25:44 ○
☐ 18 1q16.A_A1116: ○
☐ 19 1qcs.A_A59:78 ○
☐ 20 1dms_668:687 ○
☐ 21 1h0h.A_A881:9 ○
☐ 22 1tmo_677:696 ○
☐ 23 2nap.A_A642:6 ○
☐ 24 1e32.A_A38:57 ○
☐ 25 1e32.A_A81:10 ○
☐ 26 1fr3.L_L35:54 ○
☐ 27 1h9k.A_A36:55 ○
☐ 28 1h9k.A_A108:1 ○
☐ 29 1mvf.D_D14:33 ○

enter RASMOL like command...

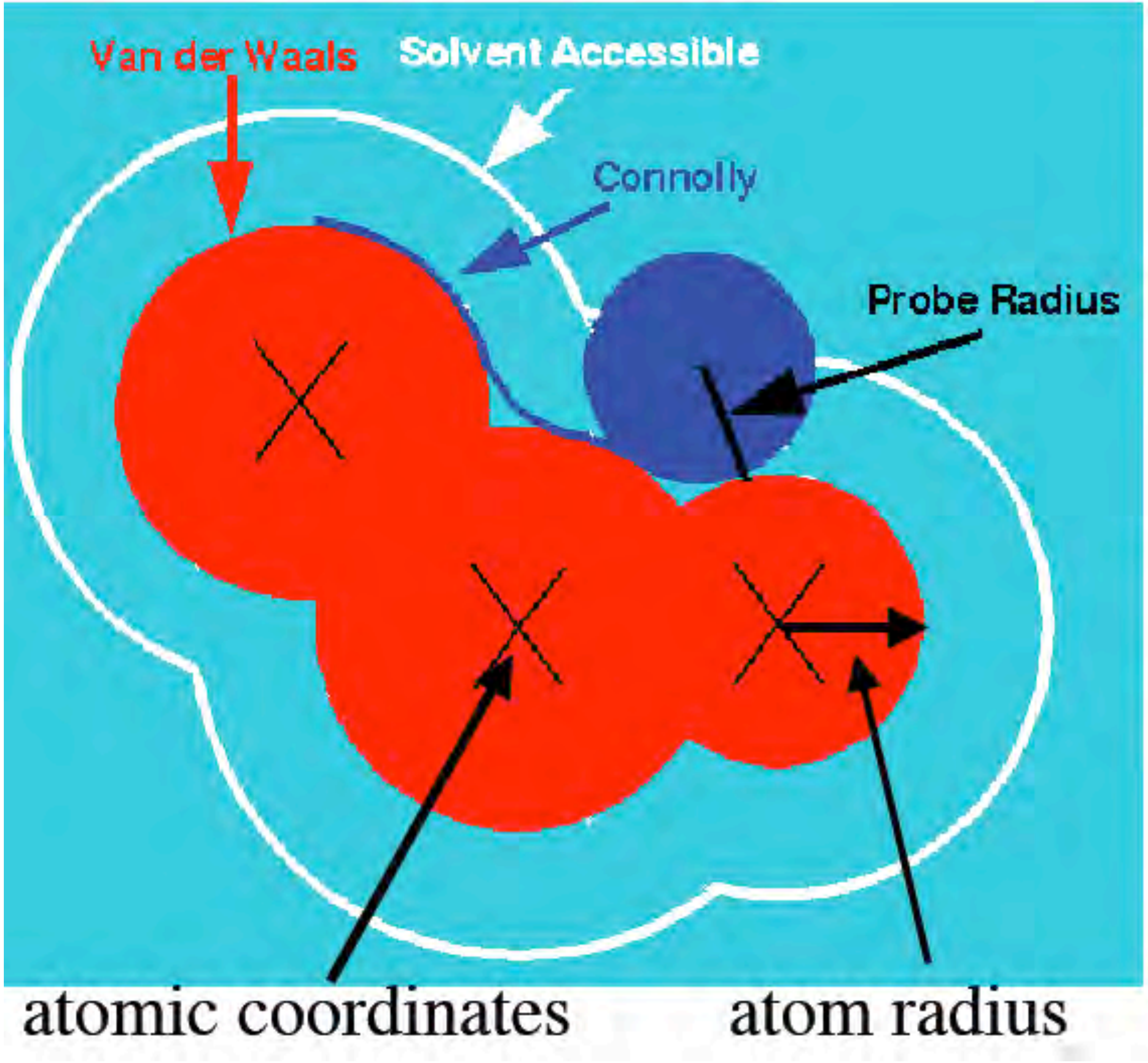
Feature: method: TIGRFAMs type: TIGRFAMsnote: Description: lp_hng_hel_AbrB Status: Tr...

Acessibilité au solvant

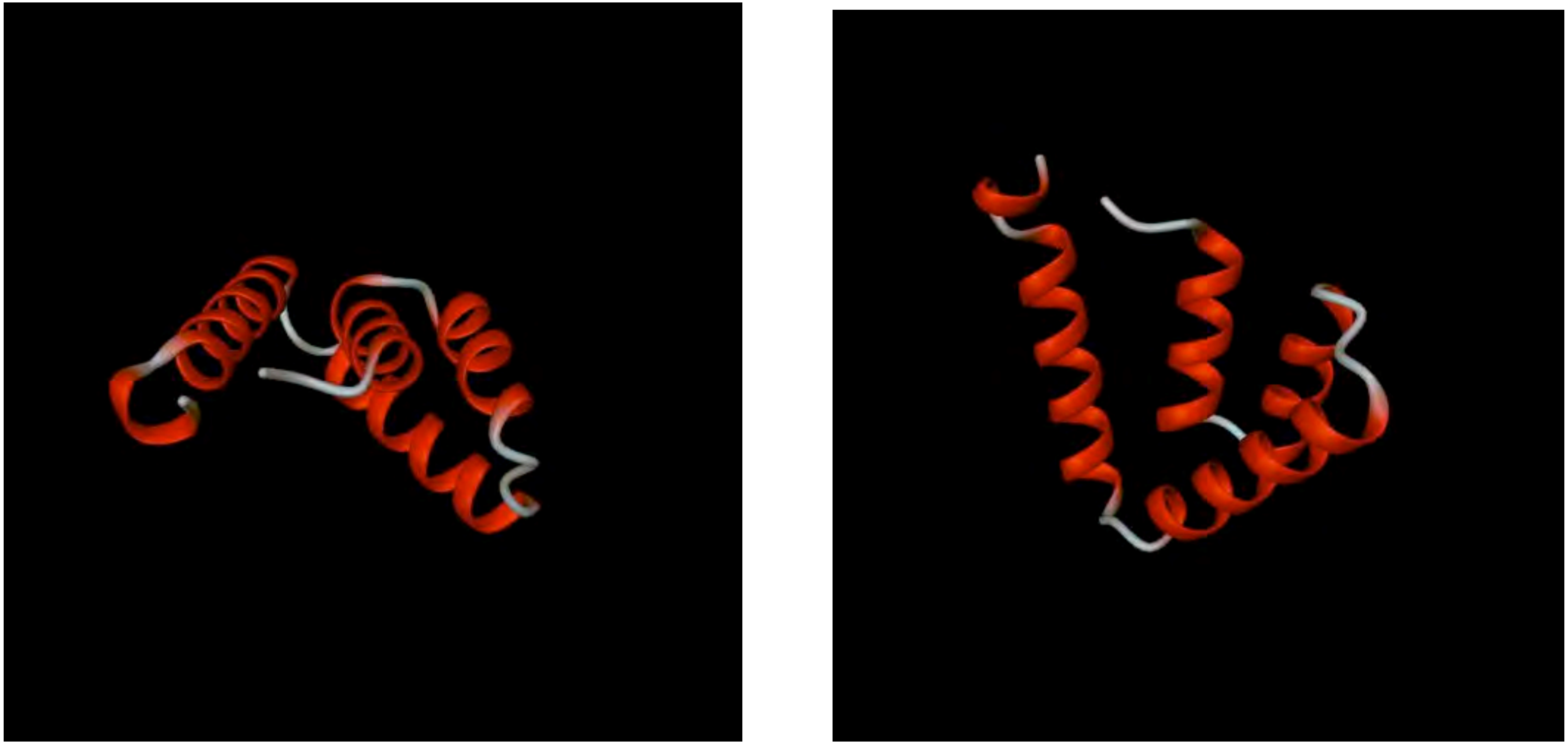


- Une sonde roule au contact des rayons de van der Waals des atomes de la protéine: elle décrit ainsi la surface moléculaire
- Le lieu du centre de la sphère sonde est la surface accessible au solvant

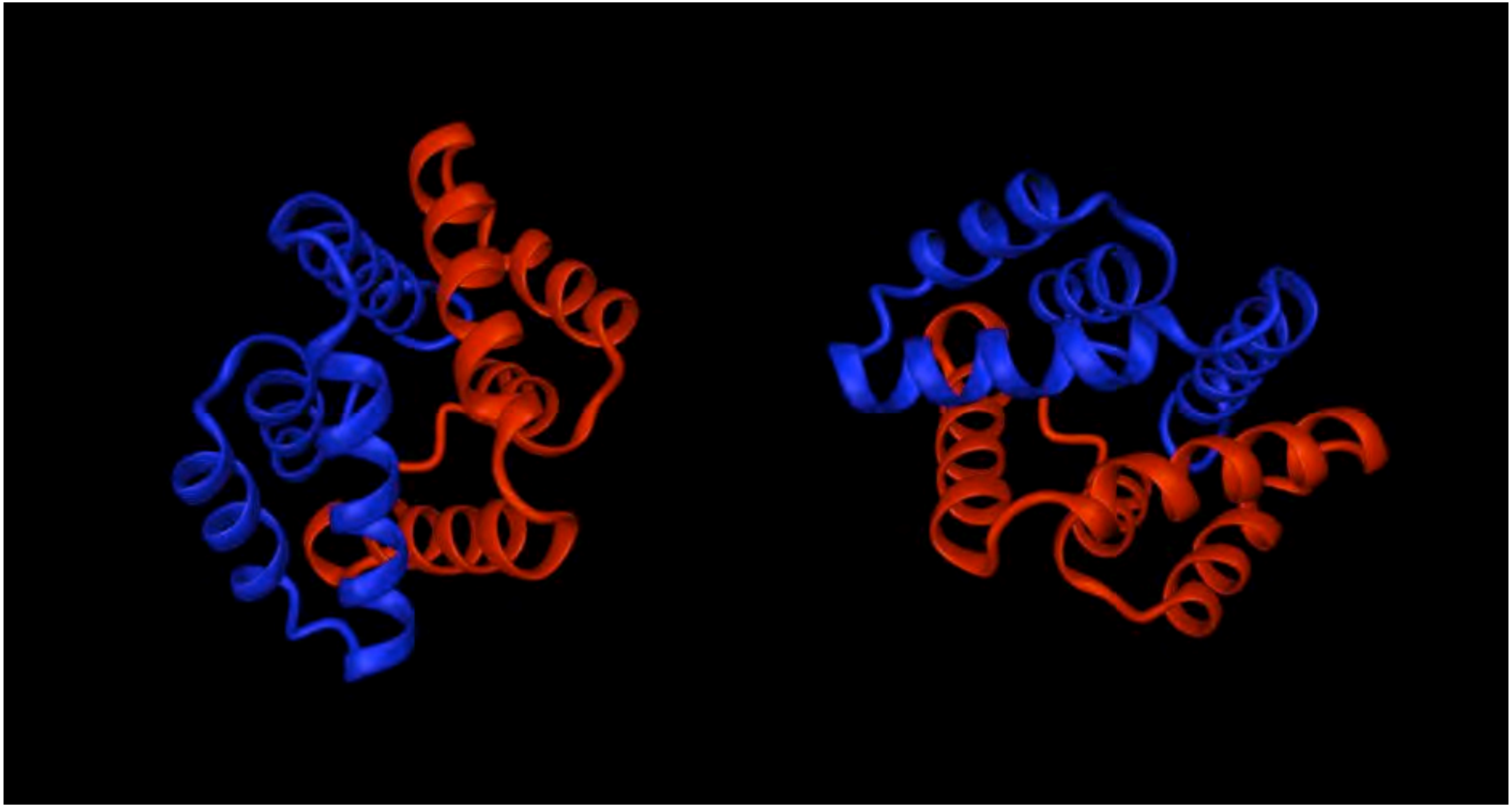
Accessibilité au solvant



1utg Monomère

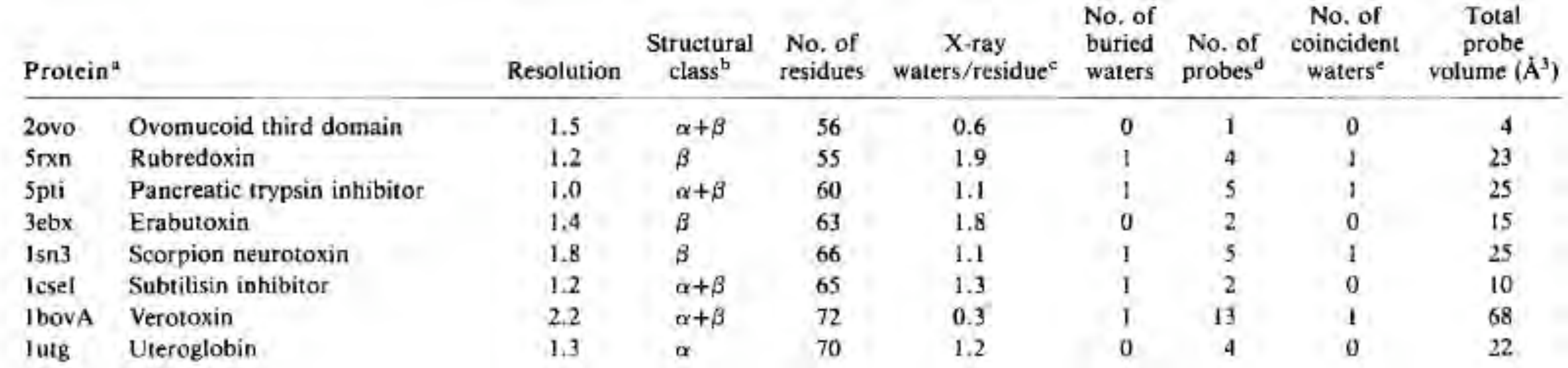


2utg Dimère

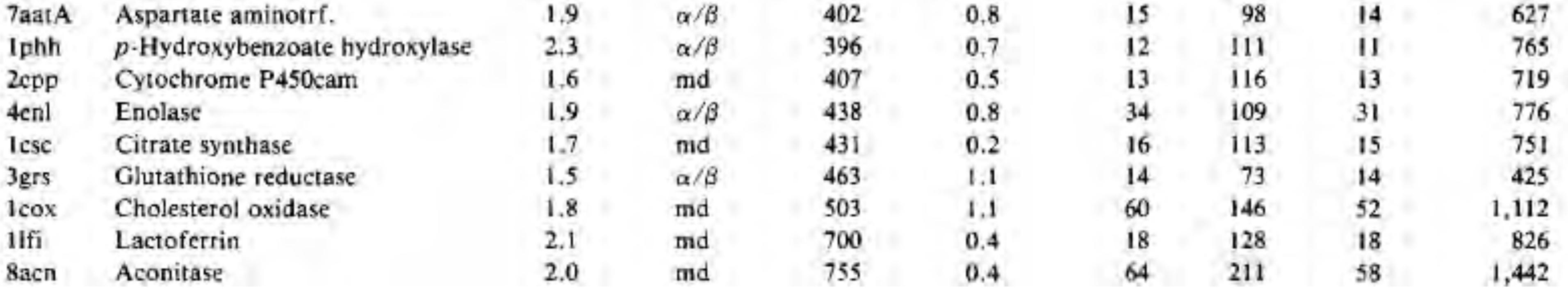


Cavités dans les protéines

Prot Sci (1994) 3:1224



Protein[a]		Resolution	Structural class[b]	No. of residues	X-ray waters/residue[c]	No. of buried waters	No. of probes[d]	No. of coincident waters[e]	Total probe volume (Å^3)
2ovo	Ovomucoid third domain	1.5	$\alpha+\beta$	56	0.6	0	1	0	4
5rxn	Rubredoxin	1.2	β	55	1.9	1	4	1	23
5pti	Pancreatic trypsin inhibitor	1.0	$\alpha+\beta$	60	1.1	1	5	1	25
3ebx	Erabutoxin	1.4	β	63	1.8	0	2	0	15
1sn3	Scorpion neurotoxin	1.8	β	66	1.1	1	5	1	25
1cseI	Subtilisin inhibitor	1.2	$\alpha+\beta$	65	1.3	1	2	0	10
1bovA	Verotoxin	2.2	$\alpha+\beta$	72	0.3	1	13	1	68
1utg	Uteroglobin	1.3	α	70	1.2	0	4	0	22



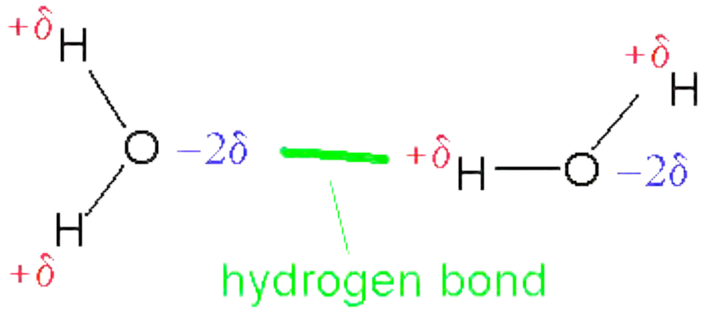
7aatA	Aspartate aminotrf.	1.9	α/β	402	0.8	15	98	14	627
1phh	*p*-Hydroxybenzoate hydroxylase	2.3	α/β	396	0.7	12	111	11	765
2cpp	Cytochrome P450cam	1.6	md	407	0.5	13	116	13	719
4enl	Enolase	1.9	α/β	438	0.8	34	109	31	776
1csc	Citrate synthase	1.7	md	431	0.2	16	113	15	751
3grs	Glutathione reductase	1.5	α/β	463	1.1	14	73	14	425
1cox	Cholesterol oxidase	1.8	md	503	1.1	60	146	52	1,112
1lfi	Lactoferrin	2.1	md	700	0.4	18	128	18	826
8acn	Aconitase	2.0	md	755	0.4	64	211	58	1,442

Attribution SS2

Attribution structures secondaires

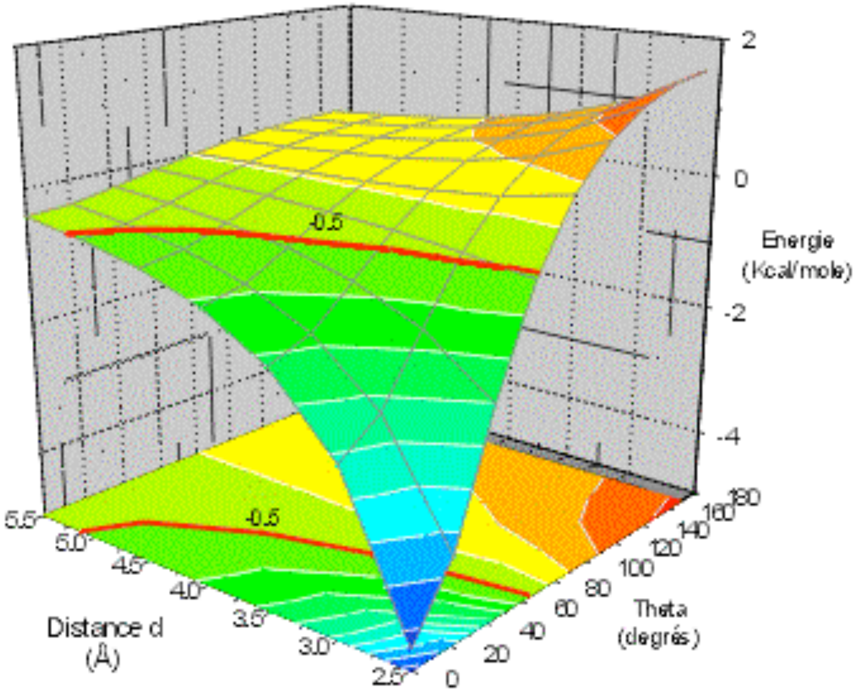
- A partir du fichier PDB; les seuls angles dièdres insuffisants
- DSSP calcule les liaisons H possibles. Si liaisons tous les (i+3), (i+4) ou (i+5) alors hélice. Si liaisons entre les atomes NiH-Oj et Oi-Nj+2 alors brins parallèles. Si liaisons entre les atomes Oi-Nj et Ni-Oj, alors brins antiparallèles.
- DSSP attribue une structure secondaire si 4 résidus successifs ont même attribution.
- DSSP attribue les tours.

DSSP

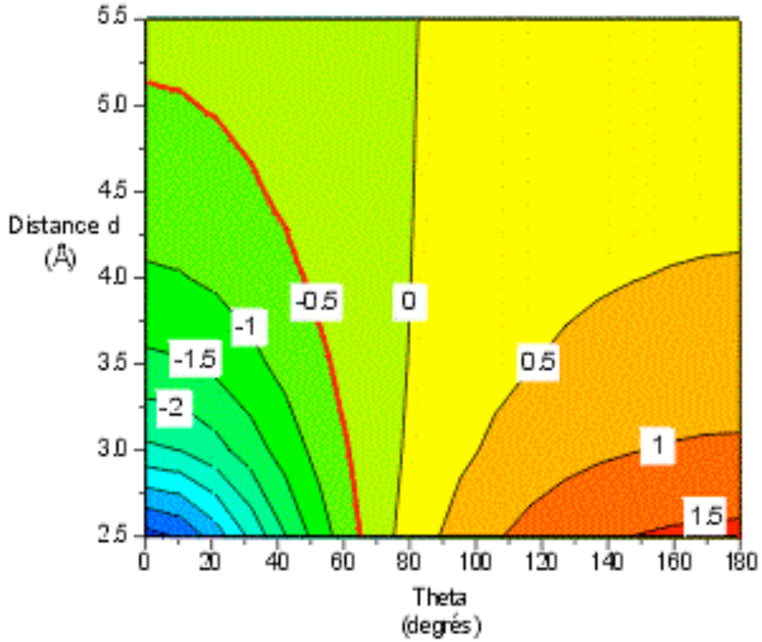


- Liaison H dans l’eau
- Energie de la liaison H en fonction de l’angle HNO et de la distance NO
- Distance canonique DA = 2.9 Å (D donneur; A accepteur)
- Angle DHA canonique 180°

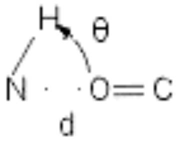
DSSP



2
0
-2
-4
Energie
(Kcal/mole)
-0.5
-0.5
5.5
5.0
4.5
4.0
3.5
3.0
2.5
Distance d
(Å)
0
20
40
60
80
100
120
140
160
180
Theta
(degrés)



5.5
5.0
4.5
4.0
3.5
3.0
2.5
Distance d
(Å)
-2
-1.5
-1
-0.5
0
0.5
1
1.5
0
20
40
60
80
100
120
140
160
180
Theta
(degrés)



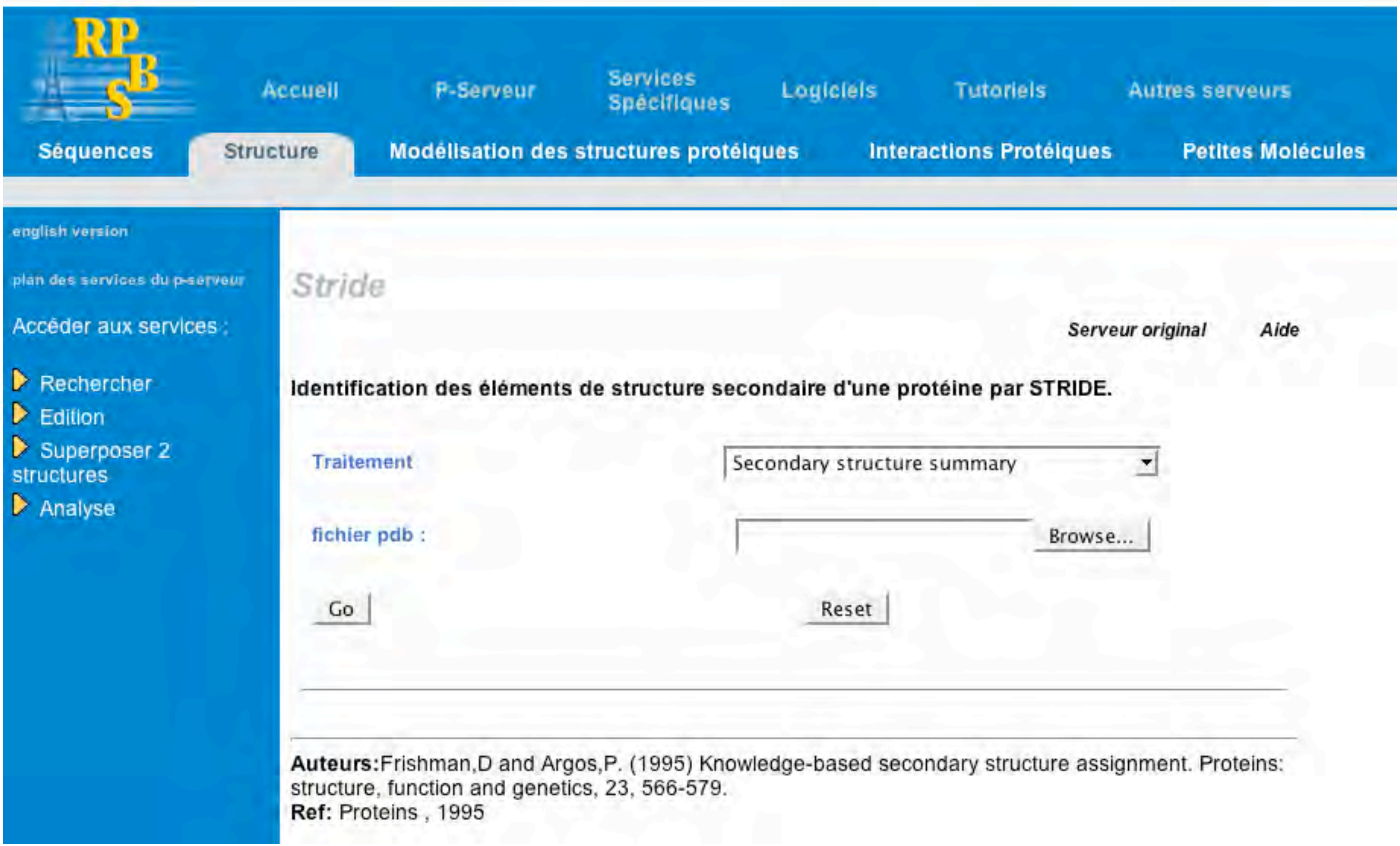
H
θ
N
O═C
d

STRIDE

Voisin de DSSP

Utilise les liaisons hydrogènes et les angles dièdres

Stride @ RPBS



Stride

Identification des éléments de structure secondaire d'une protéine par STRIDE.

Traitement: Secondary structure summary

fichier pdb :

Auteurs: Frishman,D and Argos,P. (1995) Knowledge-based secondary structure assignment. Proteins: structure, function and genetics, 23, 566-579.
Ref: Proteins , 1995

PSEA

- Utilise les distances entre carbones alpha : d2(i-1,i+2), d3 (i-1,i+3) et d4 (i,i+4)
- Utilise les pseudo angles de valence τ entre (i-1,i) et (i,i+1)
- Utilise les pseudo angles dièdres α (i-1,i,i+1,i+2)
- Comparaison à des valeurs canoniques

- Se contente des carbones alpha

P-SEA

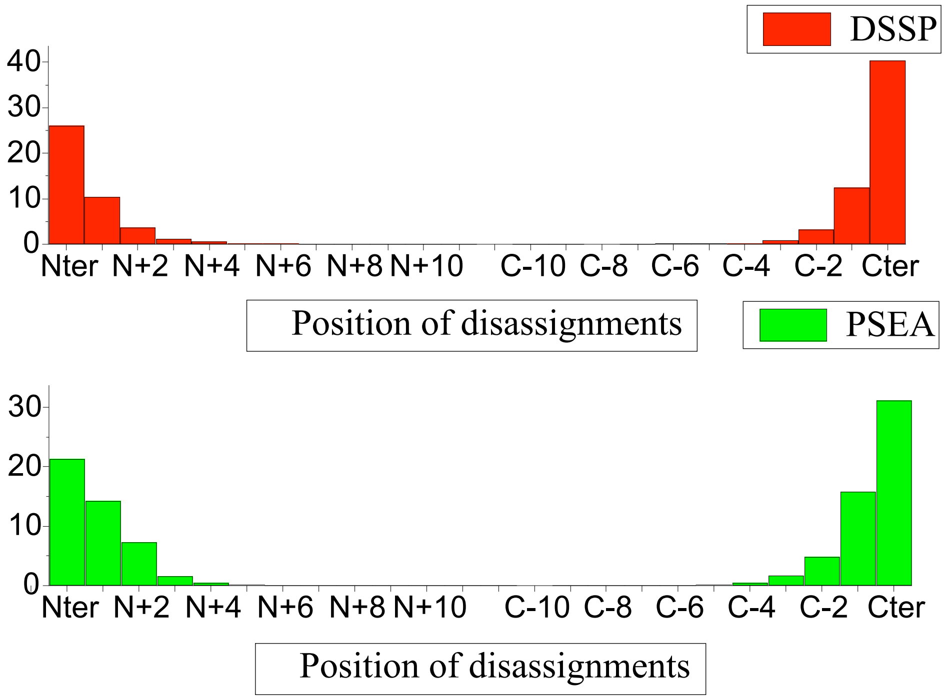
Paramètres d 'assignation des structures secondaires

Paramètre	Hélice α	Brin β
angle τ (°) dièdre α (°)	89±12 50±20	124 ±14 -170 ±45
distance d3 (Å) distance d4 (Å)	5,3 ±0,5 6,4 ±0,6	9,9 ±0,9 12,4 ±1,2

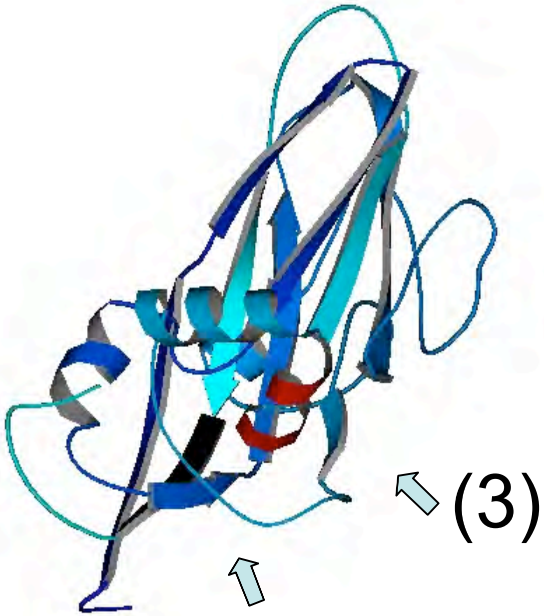
Fiabilité des attributions

- Banque de chaînes de séquences identiques (au moins des paires)
- Attribution des structures secondaires sur ces chaînes
- Peu de désaccords a/b
- Mésattributions ac ou bc essentiellement localisées aux extrémités des structures secondaires

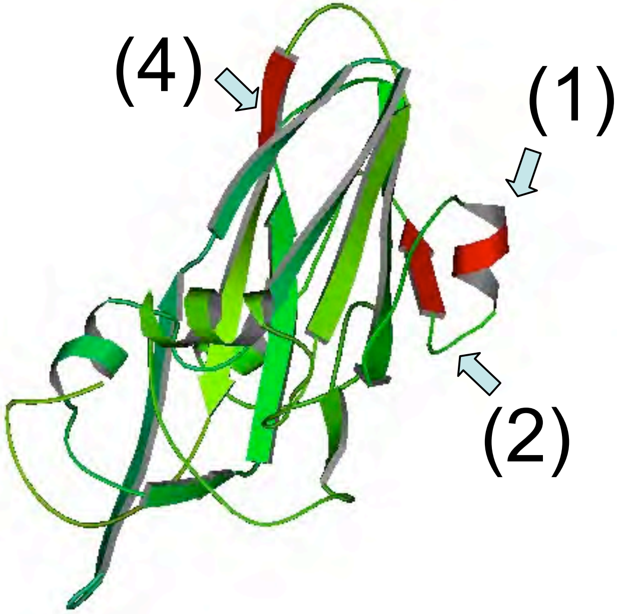
Mésattributions



- Localisation des mésattributions dans l’ensemble des structures secondaires régulières
- Localisation majoritaire aux extrémités
- Deux programmes DSSP et PSEA

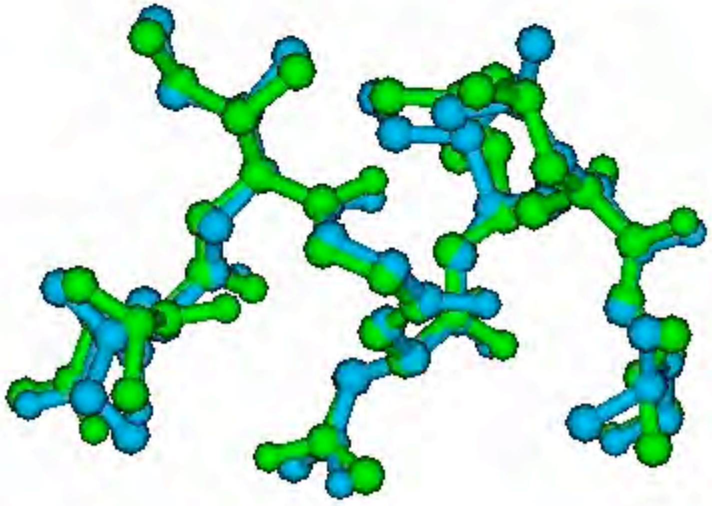


Chaîne A



Chaîne B

1) 6AA alpha
2) 4 AA beta
3) 8 AA alpha
4) brin décalé



4sbv Deux chaînes identiques à 100%
Superposition des carbones alpha 0,61 Å
rms hélice 213-220 = 0.26 Å

Hélice 213-220