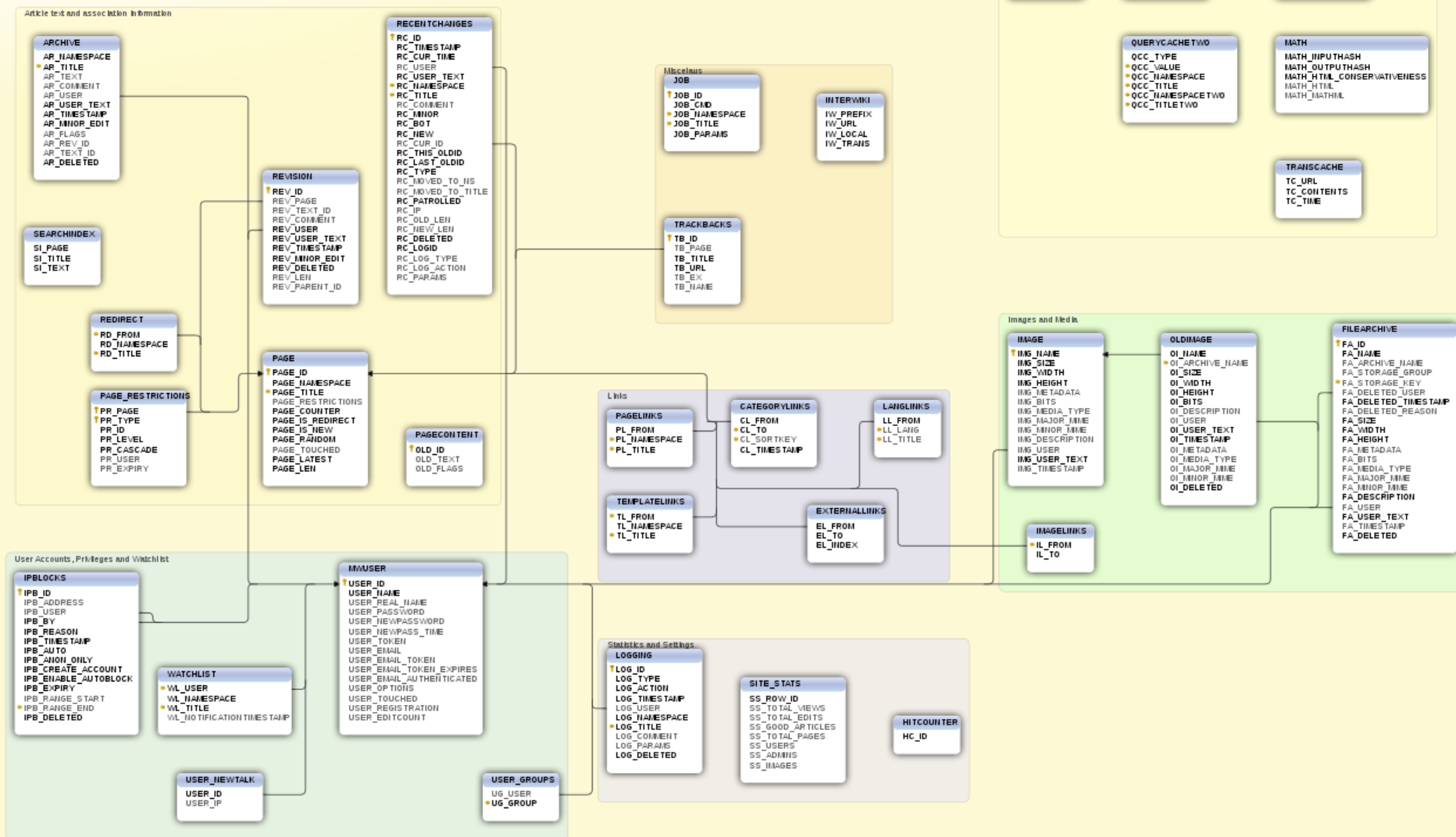


# Banques et bases de données

- Banques de données généralistes
- Banques et/ou bases de données spécialisées
- Banques de « connaissances » et autres ressources

# Exemple d'un schéma de base de données

"Media Wiki" database schema  
Move the mouse cursor over the table header and columns to view the comments  
generated us by DbSchema



# Banques généralistes

Banques de séquences d'acides nucléiques (créées dans les années 1980):

Contiennent **toutes** les séquences d'acides nucléiques produites dans les laboratoires publiques (>1000 milliards de bp, >230 millions de séquences)

Pour qu'une publication faisant référence à une ou des séquences soit acceptée, il faut que la (les) séquences ait(ent) été déposée(s) au préalable dans une de ces banques et ait(ent) obtenu un **numéro d'accension**

EMBL : la banque européenne maintenue à l'EBI (European Bioinformatics Institut) à Cambridge (UK)

GenBank : la banque américaine maintenue au NCBI (National Center for Biotechnology Information) à Bethesda (USA)

DDBJ : la banque japonaise (DNA Data Bank of Japan)

Synchronisation régulière entre ces 3 banques : INSD (International Nucleotide Sequence Database collaboration)

# Exemple d'entrée dans la banque EMBL

```
ID  AY115493; SV 1; linear; genomic DNA; STD; MUS; 48787 BP.
XX
AC  AY115493;
XX
DT  03-JUL-2002 (Rel. 72, Created)
DT  03-JUL-2002 (Rel. 72, Last updated, Version 1)
XX
DE  Mus musculus transmembrane glycoprotein E11 (E11) gene, promoter region,
DE  exons 1 through 6 and complete cds.
XX
KW  .
XX
OS  Mus musculus (house mouse)
OC  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC  Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC  Muridae; Murinae; Mus.
XX
RN  [1]
RP  1-48787
RA  Lu Y., Zhang J., Harris M.A., Harris S.E., Bonewald L., Feng J.;
RT  "Cloning and characterization studies of mouse E11 gene and its spatial and
RT  temporal expression pattern during development";
RL  Unpublished.
XX
RN  [2]
RP  1-48787
RA  Lu Y., Zhang J., Harris M.A., Harris S.E., Bonewald L., Feng J.;
RT  ;
RL  Submitted (28-MAY-2002) to the EMBL/GenBank/DDBJ databases.
RL  Oral Biology, School of Dentistry, University of Missouri-Kansas City, 650
RL  E. 25th Street, Kansas City, MO 64108, USA
XX
```

```

FH Key Location/Qualifiers
FH
FH source 1..48787
FH /organism="Mus musculus"
FH /strain="129/SV"
FH /mol_type="genomic DNA"
FH /db_xref="taxon:10090"
FH promoter 1..9640
FH /gene="E11"
FH mRNA join(9641..9915,33318..33454,35464..35620,39063..39101,
FH 40324..40435,41199..42193)
FH /gene="E11"
FH /product="transmembrane glycoprotein E11"
FH exon 9641..9915
FH /gene="E11"
FH /number=1
FH CDS join(9849..9915,33318..33454,35464..35620,39063..39101,
FH 40324..40435,41199..42193)
FH /codon_start=1
FH /gene="E11"
FH /product="transmembrane glycoprotein E11"
FH /note="gp38; PA2.26; OTS8"
FH /db_xref="GOA:Q62011"
FH /db_xref="InterPro:IPR008783"
FH /db_xref="MGI:103098"
FH /db_xref="UniProtKB/Swiss-Prot:Q62011"
FH /protein_id="AAM66761.1"
FH /translation="MWTVPVLFWVLGVSVWFWDSAQGGTIGVNEDDIVTPGTGDGMVPPG
FH IEDKITTTGATGGLNESTGKAPLVPTQRRERGTPKPPLEELSTSATSDHDHREHESTTTVK
FH VVTSHSVDKKTSHPNRDNAGDETTQTDKKDGLPVVTLVGIIVGVLLAIGFVGGIFIVVM
FH KKISGRFSP"
FH exon 33318..33454
FH /gene="E11"
FH /number=2
FH exon 35464..35620
FH /gene="E11"
FH /number=3
FH exon 39063..39101
FH /gene="E11"
FH /number=4
FH exon 40324..40435
FH /gene="E11"
FH /number=5
FH exon 41199..42193
FH /gene="E11"
FH /number=6
FH polyA_signal 42172..42177
FH /gene="E11"

```

## Exemple d'entrée dans la banque EMBL (suite)

# Exemple d'entrée dans la banque EMBL

```
SQ   Sequence 48787 BP; 12689 A; 11292 C; 11677 G; 13094 T; 35 other;
      tatagtcact cttgccaaatt gcactaagga taacccaacc ttggtttaaa aaaaaaaatc           60
      ctgactcaaa gatgaatttc acacttagca gagacattta cctgccaaaga aagacacaat           120
      gcacccccat gtcagctccc ctccacctca ccccatgaca ccccaaagcc tgtggtgctt           180
      tcttaagaga ggctgttcga acaccgtgtc ctactgcttt ccactgccag gacgacaggc           240
      aattcttcag cctaggtaaa cttcaggaaa ataaagttta atcaagagct gtgtctggtg           300
      acggctttct taacatgaga atccaaaggg gtagctatat gacctggaaa aagacagtga           360
      ttcaggggtg tacacgtggg tgtgtacttg cctctgtgtg catctgcgtg tgtgtatgtc           420
      tctctgtgtg tgcacacccg agtggcgaat gtttgtatgt agtgtgtgta tgtatgtgtg           480
      tgtgtctgtg tgctgtgtgg tatatgtgtg tgcgcgcgtg agagtgtgta atgtgtctgg           540
      gctctccttg ttataaaatc accaggatcc cactatgaca cgcacatacg aactgccttc           600
      caaaggctcc acctctgaat gacacaaccc cgtcgccttat cctctcagca cctcaaagag           660
      aagattaatt tttcaacaca ggaatccttt ggggacgact tccaaaccgc agcacacagc           720
      aaccctgaat gaaatctgca cgtctggggg caacgctcca ccttaggagc aagcatagct           780
      gagggctttg tgtttctgta tcaaagtcca aagtggaaaa agaacagagg acgggggaag           840
      .....
      gcaaagctct aggtcaatga gaaaccctgt ctcaaacaaa agggagaagc ccctaaggcc           48600
      tgacagctgg ggtttgtcct ctgatttcca tgcgcatgag cacggatatg cacacatata           48660
      cctgcataca cacacacaca cacacacaca cacacacaca agcatactcg tgcacatatg           48720
      tgcattcata taaacacata cacacaaaaa tgaaccttat cttatttaat tacttttttt           48780
      ggcacag                                           48787
```

//

# Exemple d'entrée dans la banque GenBank

```
LOCUS          AY115493          48787 bp      DNA      linear      ROD 05-JUN-2006
DEFINITION    Mus musculus transmembrane glycoprotein E11 (E11) gene, promoter
              region, exons 1 through 6 and complete cds.
ACCESSION     AY115493
VERSION       AY115493.1  GI:21684686
KEYWORDS      .
SOURCE        Mus musculus (house mouse)
  ORGANISM    Mus musculus
              Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
              Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
              Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE     1 (bases 1 to 48787)
  AUTHORS     Zhang,K., Barragan-Adjemian,C., Ye,L., Kotha,S., Dallas,M., Lu,Y.,
              Zhao,S., Harris,M., Harris,S.E., Feng,J.Q. and Bonewald,L.F.
  TITLE       E11/gp38 Selective Expression in Osteocytes: Regulation by
              Mechanical Strain and Role in Dendrite Elongation
  JOURNAL     Mol. Cell. Biol. 26 (12), 4539-4552 (2006)
  PUBMED     16738320
REFERENCE     2 (bases 1 to 48787)
  AUTHORS     Lu,Y., Zhang,J., Harris,M.A., Harris,S.E., Bonewald,L. and Feng,J.
  TITLE       Cloning and characterization studies of mouse E11 gene and its
              spatial and temporal expression pattern during development
  JOURNAL     Unpublished
REFERENCE     3 (bases 1 to 48787)
  AUTHORS     Lu,Y., Zhang,J., Harris,M.A., Harris,S.E., Bonewald,L. and Feng,J.
  TITLE       Direct Submission
  JOURNAL     Submitted (28-MAY-2002) Oral Biology, School of Dentistry,
              University of Missouri-Kansas City, 650 E. 25th Street, Kansas
              City, MO 64108, USA
```

| FEATURES     | Location/Qualifiers   |
|--------------|---|
| source       | 1..48787<br>/organism="Mus musculus"<br>/mol_type="genomic DNA"<br>/strain="129/Sv"<br>/db_xref="taxon:10090"   |
| gene         | 1..42193<br>/gene="E11"   |
| promoter     | 1..9640<br>/gene="E11"  |
| mRNA         | join(9641..9915,33318..33454,35464..35620,39063..39101,<br>40324..40435,41199..42193)<br>/gene="E11"<br>/product="transmembrane glycoprotein E11"   |
| exon         | 9641..9915<br>/gene="E11"<br>/number=1  |
| CDS          | join(9849..9915,33318..33454,35464..35620,39063..39101,<br>40324..40435,41199..41205)<br>/gene="E11"<br>/note="gp38; PA2.26; OTS8"<br>/codon_start=1<br>/product="transmembrane glycoprotein E11"<br>/protein_id="AAM66761.1"<br>/db_xref="GI:21684687"<br>/translation="MWTVPVLFVWLGSVFWWSAQGGTIGVNEDDIVTPGTGDMVPP<br>GIEDKITTGTGATGGLNESTGKAPLVPTQRRERGTKPPLLEELSTSATSDDHREHESTTT<br>VKVVTSHSVDKKTSHPNRDNAGDETTQTTDKKDGLPVVTLVGIIVGVLLAIGFVGGIFI<br>VVMKKISGRFSP" |
| exon         | 33318..33454<br>/gene="E11"<br>/number=2  |
| exon         | 35464..35620<br>/gene="E11"<br>/number=3  |
| exon         | 39063..39101<br>/gene="E11"<br>/number=4  |
| exon         | 40324..40435<br>/gene="E11"<br>/number=5  |
| exon         | 41199..42193<br>/gene="E11"<br>/number=6  |
| polyA_signal | 42172..42177<br>/gene="E11"   |

Exemple d'entrée dans la banque GenBank (suite)



# Banques généralistes

## Banques de séquences protéiques :

Les deux plus importantes :

SwissProt (1986) : banque manuellement annotée et « nettoyée »

PIR/NBRF (1984) : banque américaine fournissant une classification des protéines basée sur la similarité entre les séquences.

TrEMBL : traduction automatique des CDS d'EMBL

GenPept : traduction automatique des CDS de GenBank

En 2002, création du consortium **UniProt** (Universal Protein Resource) constitué par le groupe SwissProt-TrEMBL et le groupe PIR

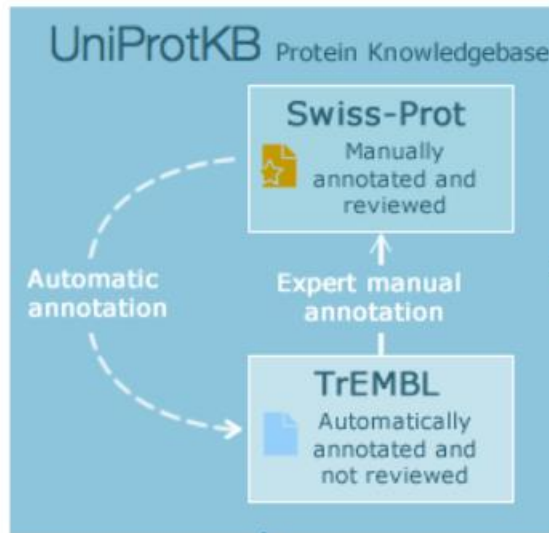
But : fournir une seule ressource centralisée pour les séquences protéiques et les annotations fonctionnelles

Maintien de deux sections enrichies de la classification automatique de PIR : **UniProt/SwissProt** (annotée et « nettoyée »)

**UniProt/TrEMBL**

UniProt aujourd'hui: ~ 225 millions de séquences (dont 566 000 de SwissProt)

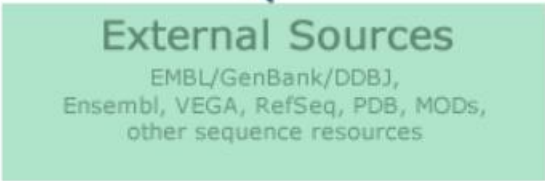
# UniProtKB



Séquences protéiques des génomes complètement séquencés



Séquences protéiques provenant de UniProtKB et UniParc regroupées en clusters en fonction de leur pourcentage d'identité (100%, 90%, 50%).



UniParc est une base de données non redondante contenant la majorité des séquences protéiques disponibles dans le monde. Des liens vers d'autres bases de données permettent d'accéder à des informations complémentaires.

# Exemple d'entrée dans la banque SwissProt

ID PDPN\_MOUSE STANDARD; PRT; 172 AA.  
AC Q62011; Q546R8; Q61612;  
DT 01-NOV-1997, integrated into UniProtKB/Swiss-Prot.  
DT 01-NOV-1997, sequence version 2.  
DT 19-SEP-2006, entry version 45.  
DE Podoplanin precursor (Glycoprotein 38) (Gp38) (OTS-8) (PA2.26 antigen)  
DE (Aggrus) (T1A) (T1-alpha) (Transmembrane glycoprotein E11).  
GN Name=Pdpn; Synonyms=Gp38, Ots8;  
OS Mus musculus (Mouse).  
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
OC Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi;  
OC Muroidea; Muridae; Murinae; Mus.  
OX NCBI\_TaxID=10090;  
RN [1]  
RP NUCLEOTIDE SEQUENCE [MRNA].  
RX MEDLINE=91207913; PubMed=2088477;  
RA Nose K., Saito H., Kuroki T.;  
RT "Isolation of a gene sequence induced later by tumor-promoting 12-O-  
RT tetradecanoylphorbol-13-acetate in mouse osteoblastic cells (MC3T3-E1)  
RT and expressed constitutively in ras-transformed cells.";  
RL Cell Growth Differ. 1:511-518(1990).  
RN [2]  
RP NUCLEOTIDE SEQUENCE [MRNA].  
RC STRAIN=BALB/c;  
RX MEDLINE=93018879; PubMed=1402691; DOI=10.1084/jem.176.5.1477;  
RA Farr A.G., Berry M.L., Kim A., Nelson A.J., Welch M.P., Aruffo A.;  
RT "Characterization and cloning of a novel glycoprotein expressed by  
RT stromal cells in T-dependent areas of peripheral lymphoid tissues.";  
RL J. Exp. Med. 176:1477-1482(1992).  
RN [3]  
RP NUCLEOTIDE SEQUENCE [MRNA], PROTEIN SEQUENCE OF 24-30 AND 66-73,  
RP FUNCTION, SUBCELLULAR LOCATION, TISSUE SPECIFICITY, AND GLYCOSYLATION.  
RX MEDLINE=20044810; PubMed=10574709;  
RA Scholl F.G., Gamallo C., Vilaro S., Quintanilla M.;  
RT "Identification of PA2.26 antigen as a novel cell-surface mucin-type  
RT glycoprotein that induces plasma membrane extensions and increased  
RT motility in keratinocytes.";  
RL J. Cell Sci. 112:4601-4613(1999)

# Exemple d'entrée dans la banque SwissProt (suite)

```
CC      -!- FUNCTION: May be involved in cell migration and/or actin
CC      cytoskeleton organization. When expressed in keratinocytes,
CC      induces changes in cell morphology with transfected cells showing
CC      an elongated shape, numerous membrane protrusions, major
CC      reorganization of the actin cytoskeleton, increased motility and
CC      decreased cell adhesion. Required for normal lung cell
CC      proliferation and alveolus formation at birth. Induces platelet
CC      aggregation. Does not have any effect on folic acid or amino acid
CC      transport. Does not function as a water channel or as a regulator
CC      of aquaporin-type water channels.
CC      -!- SUBCELLULAR LOCATION: Membrane; single-pass type I membrane
CC      protein. Localized to actin-rich microvilli and plasma membrane
CC      projections such as filopodia, lamellipodia and ruffles.
CC      -!- TISSUE SPECIFICITY: Detected at high levels in lung and brain, at
CC      lower levels in kidney, stomach, liver, spleen and esophagus, and
CC      not detected in skin and small intestine. Expressed in epithelial
CC      cells of choroid plexus, ependyma, glomerulus and alveolus, in
CC      mesothelial cells and in endothelia of lymphatic vessels. Also
CC      expressed in stromal cells of peripheral lymphoid tissue and
CC      thymic epithelial cells. Detected in carcinoma cell lines and
CC      cultured fibroblasts. Expressed at higher levels in colon
CC      carcinomas than in normal colon tissue.
CC      -!- INDUCTION: Down-regulated by treatment with puromycin
CC      aminonucleoside.
CC      -!- PTM: Extensively O-glycosylated. Contains sialic acid residues. O-
CC      glycosylation is necessary for platelet aggregation activity.
CC      -!- PTM: The N-terminus is blocked (By similarity).
CC      -!- MISCELLANEOUS: Mice lacking Pdpn die at birth of respiratory
CC      failure due to a low number of attenuated type I cells, narrow and
CC      irregular air spaces, and defective formation of alveolar
CC      saccules.
CC      -!- SIMILARITY: Belongs to the podoplanin family.
CC      -----
CC      Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC      Distributed under the Creative Commons Attribution-NoDerivs License
CC      -----
```

# Exemple d'entrée dans la banque SwissProt (suite)

```
DR   EMBL; M73748; AAA39866.1; -; mRNA.
DR   EMBL; M96645; AAA37724.1; -; mRNA.
DR   EMBL; AJ250246; CAB58997.1; -; mRNA.
DR   EMBL; AJ297944; CAC16152.1; -; mRNA.
DR   EMBL; AY115493; AAM66761.1; -; Genomic_DNA.
DR   EMBL; AK158855; BAE34695.1; -; mRNA.
DR   EMBL; BC026551; AAH26551.1; -; mRNA.
DR   Ensembl; ENSMUSG00000028583; Mus musculus.
DR   KEGG; mmu:14726; -.
DR   MGI; MGI:103098; Pdpn.
DR   ArrayExpress; Q62011; -.
DR   RZPD-ProtExp; IOM20239; -.
DR   GO; GO:0030175; C:filopodium; IDA.
DR   GO; GO:0030027; C:lamellipodium; IDA.
DR   GO; GO:0005886; C:plasma membrane; IDA.
DR   GO; GO:0001726; C:ruffle; IDA.
DR   GO; GO:0000902; P:cellular morphogenesis; IDA.
DR   GO; GO:0030324; P:lung development; IMP.
DR   GO; GO:0001946; P:lymphangiogenesis; IMP.
DR   GO; GO:0051272; P:positive regulation of cell motility; IDA.
DR   InterPro; IPR008783; Podoplanin.
DR   PANTHER; PTHR16861; Podoplanin; 1.
DR   Pfam; PF05808; Podoplanin; 1.
KW   Cell shape; Developmental protein; Direct protein sequencing;
KW   Glycoprotein; Membrane; Sialic acid; Signal; Transmembrane.
FT   SIGNAL          1      22      Potential.
FT   CHAIN           23     172     Podoplanin.
FT                                     /FTId=PRO_0000021352.
FT   TOPO_DOM       23     141     Extracellular (Potential).
FT   TRANSMEM      142     162     Potential.
FT   TOPO_DOM       163     172     Cytoplasmic.
.....
FT   CONFLICT       29     31      EDD -> KNN (in Ref. 2).
FT   CONFLICT       38     39      GD -> EN (in Ref. 1).
SQ   SEQUENCE       172 AA;  18233 MW;  C035ED251918CE6F CRC64;
      MWTVPVLFVW LGSVWFWDSA QGGTIGVNED DIVTPGTGDG MVPPGIEDKI TTTGATGGLN
      ESTGKAPLVP TQRRERGTKPP LEELSTSATS DHDHREHEST TTVKVVTSHS VDKKTSHPNR
      DNAGDETQTT DKKDGLPVVT LVGIIVGVLL AIGFVGGIFI VVMKKISGRF SP
```

//

# Banques généralistes

## Banques de structures :

La Protein Database (PDB) stockent les structures protéiques obtenues par RMN ou cristallographie

Une entrée contient donc les coordonnées de tous les atomes de la structure

# Exemple d'entrée dans la banque PDB

```
HEADER PERIPLASMIC BINDING PROTEIN 17-AUG-97 4MBP
TITLE MALTODEXTRIN BINDING PROTEIN WITH BOUND MALTETROSE
COMPND MOL_ID: 1;
COMPND 2 MOLECULE: MALTODEXTRIN BINDING PROTEIN;
COMPND 3 CHAIN: NULL
SOURCE MOL_ID: 1;
SOURCE 2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI;
SOURCE 3 STRAIN: K12;
SOURCE 4 CELLULAR_LOCATION: PERIPLASM;
SOURCE 5 GENE: MALE
KEYWDS PERIPLASMIC BINDING PROTEIN, TRANSPORT, SUGAR TRANSPORT
EXPDTA X-RAY DIFFRACTION
AUTHOR J.C.SPURLINO,F.A.QUIOCHO
REVDAT 1 25-FEB-98 4MBP 0
JRNL AUTH F.A.QUIOCHO,J.C.SPURLINO,L.E.RODSETH
JRNL TITL EXTENSIVE FEATURES OF TIGHT OLIGOSACCHARIDE BINDING
JRNL TITL 2 REVEALED IN HIGH-RESOLUTION STRUCTURES OF THE
JRNL TITL 3 MALTODEXTRIN TRANSPORT/CHEMOSENSORY RECEPTOR
JRNL REF STRUCTURE (LONDON) V. 5 997 1997
JRNL REFN ASTM STRUE6 UK ISSN 0969-2126 2005
REMARK 1
REMARK 1 REFERENCE 1
REMARK 1 AUTH A.J.SHARFF,L.E.RODSETH,F.A.QUIOCHO
REMARK 1 TITL REFINED 1.8-A STRUCTURE REVEALS THE MODE OF BINDING
REMARK 1 TITL 2 OF BETA-CYCLODEXTRIN TO THE MALTODEXTRIN BINDING
REMARK 1 TITL 3 PROTEIN
REMARK 1 REF BIOCHEMISTRY V. 32 10553 1993
REMARK 1 REFN ASTM BICHAW US ISSN 0006-2960 0033
REMARK 2 RESOLUTION. 1.7 ANGSTROMS.
REMARK 3
REMARK 3 REFINEMENT.
REMARK 3 PROGRAM : PROLSQ
REMARK 3 AUTHORS : KONNERT,HENDRICKSON
REMARK 3
REMARK 3 DATA USED IN REFINEMENT.
REMARK 3 RESOLUTION RANGE HIGH (ANGSTROMS) : 1.7
REMARK 3 RESOLUTION RANGE LOW (ANGSTROMS) : 10.0
REMARK 3 DATA CUTOFF (SIGMA(F)) : 2.0
REMARK 3 COMPLETENESS FOR RANGE (%) : 89.
REMARK 3 NUMBER OF REFLECTIONS : 29814
REMARK 3
REMARK 3 FIT TO DATA USED IN REFINEMENT.
REMARK 3 CROSS-VALIDATION METHOD : NULL
REMARK 3 FREE R VALUE TEST SET SELECTION : NULL
REMARK 3 R VALUE (WORKING + TEST SET) : NULL
```

## Exemple d'entrée dans la banque PDB (suite)

```

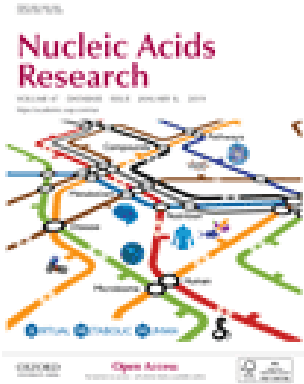
DBREF 4MBP      1  370 SWS      P02928  MALE_ECOLI      27  396
SEQRES  1      370  LYS ILE GLU  GLU GLY LYS LEU VAL ILE TRP ILE ASN GLY
.....
SEQRES  27      370  PHE TRP TYR  ALA VAL ARG THR ALA VAL ILE ASN ALA ALA
SEQRES  28      370  SER GLY ARG  GLN THR VAL ASP GLU ALA LEU LYS ASP ALA
SEQRES  29      370  GLN THR ARG  ILE THR LYS
.....
ATOM    1  N   LYS      1   -14.189  28.577  49.986  1.00  59.37      N
ATOM    2  CA  LYS      1   -14.427  27.969  48.643  1.00  60.26      C
ATOM    3  C   LYS      1   -14.388  26.456  48.756  1.00  60.10      C
ATOM    4  O   LYS      1   -14.139  25.752  47.779  1.00  60.59      O
ATOM    5  CB  LYS      1   -13.396  28.459  47.636  1.00  60.34      C
ATOM    6  N   ILE      2   -14.546  25.990  49.995  1.00  60.57      N
ATOM    7  CA  ILE      2   -14.875  24.597  50.323  1.00  59.62      C
ATOM    8  C   ILE      2   -15.574  23.742  49.238  1.00  59.40      C
ATOM    9  O   ILE      2   -15.081  22.687  48.873  1.00  60.19      O
ATOM   10  CB  ILE      2   -15.681  24.585  51.659  1.00  58.28      C
ATOM   11  CG1 ILE      2   -14.724  24.301  52.826  1.00  59.82      C
ATOM   12  CG2 ILE      2   -16.838  23.622  51.603  1.00  57.23      C
ATOM   13  CD1 ILE      2   -15.337  24.388  54.226  1.00  59.04      C
ATOM   14  N   GLU      3   -16.602  24.290  48.599  1.00  61.26      N
ATOM   15  CA  GLU      3   -17.646  23.482  47.944  1.00  61.93      C
ATOM   16  C   GLU      3   -18.288  22.453  48.893  1.00  62.35      C
ATOM   17  O   GLU      3   -19.019  22.844  49.814  1.00  62.86      O
ATOM   18  CB  GLU      3   -17.108  22.769  46.711  1.00  61.59      C
ATOM   19  CG  GLU      3   -18.212  22.251  45.809  1.00  60.99      C
ATOM   20  CD  GLU      3   -18.778  23.313  44.867  1.00  63.62      C
ATOM   21  OE1  GLU      3   -18.216  24.442  44.804  1.00  63.52      O
ATOM   22  OE2  GLU      3   -19.806  23.016  44.199  1.00  64.93      O
ATOM   23  N   GLU      4   -17.924  21.178  48.713  1.00  60.97      N
ATOM   24  CA  GLU      4   -18.411  20.010  49.472  1.00  59.71      C
ATOM   25  C   GLU      4   -19.240  19.084  48.606  1.00  57.65      C
ATOM   26  O   GLU      4   -19.990  19.524  47.740  1.00  58.14      O
ATOM   27  CB  GLU      4   -19.222  20.381  50.704  1.00  60.49      C
ATOM   28  CG  GLU      4   -20.145  19.273  51.192  1.00  66.79      C
ATOM   29  CD  GLU      4   -19.416  17.960  51.537  1.00  71.32      C
ATOM   30  OE1  GLU      4   -18.345  18.014  52.202  1.00  74.47      O
ATOM   31  OE2  GLU      4   -19.941  16.873  51.166  1.00  71.40      O
ATOM   32  N   GLY      5   -19.052  17.787  48.793  1.00  56.22      N
ATOM   33  CA  GLY      5   -19.860  16.837  48.055  1.00  56.05      C
ATOM   34  C   GLY      5   -19.171  16.369  46.800  1.00  54.46      C
ATOM   35  O   GLY      5   -19.574  15.345  46.223  1.00  54.62      O
.....

```





# Banques/bases de données spécialisées



Chaque année, en janvier, le journal Nucleic Acids Research publie un numéro spécial dédié aux bases de données

## **The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection**

Daniel J. Rigden, and Xosé M. Fernandez

“The 2019 Nucleic Acids Research (NAR) Database Issue contains 168 papers spanning molecular biology. Among them, 64 are new and another 92 are updates describing resources that appeared in the Issue previously. The remaining 12 are updates on databases most recently published elsewhere. This Issue contains two Breakthrough articles, on the Virtual Metabolic Human (VMH) database which links human and gut microbiota metabolism with diet and disease, and Vibrism DB, a database of mouse brain anatomy and gene (co-)expression with sophisticated visualization and session sharing.” .....

“The entire database issue is freely available online on the NAR website (<https://academic.oup.com/nar>). The NAR online Molecular Biology Database Collection has been updated, reviewing 506 entries, adding 66 new resources and eliminating 147 discontinued URLs, bringing the current total to 1613 databases. It is available at <http://www.oxfordjournals.org/nar/database/c>.”

# Banques/bases de données spécialisées

Certaines modélisent des chemins métaboliques ou des processus de régulation :

- Regulondb (regulation) et EcoCyc (métabolisme) pour *E. coli*
- Extension de EcoCyc à MetaCyc (multiorganismes, surtout microorganismes et plantes) et AraCyc (*Arabidopsis thaliana*)

- KEGG : Kyoto Encyclopedia of Genes and Genomes

Quatre parties :

- **Pathway** database
- Genes database
- Genome database
- Orthology database
- ...



KEGG ▾

Search

[Help](#)[» Japanese](#)

### KEGG Home

[Release notes](#)  
[Current statistics](#)  
[Plea from KEGG](#)

### KEGG Database

[KEGG overview](#)  
[Searching KEGG](#)  
[KEGG mapping](#)  
[Color codes](#)

### KEGG Objects

[Pathway maps](#)  
[Brite hierarchies](#)

### KEGG Software

[KegTools](#)  
[KEGG API](#)  
[KGML](#)

### KEGG FTP

[Subscription](#)

[GenomeNet](#)

[DBGET/LinkDB](#)

[Feedback](#)

[Kanehisa Labs](#)

## KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (See [Release notes](#) for new and updated features).

### ● Main entry point to the KEGG web service

**KEGG2** [KEGG Table of Contents](#) [Update notes](#)

### ● Data-oriented entry points

**KEGG PATHWAY** [KEGG pathway maps](#) [[Pathway list](#)]  
**KEGG BRITE** [BRITE functional hierarchies](#) [[Brite list](#)]  
**KEGG MODULE** [KEGG modules](#) [[Module list](#) | [Statistics](#)]  
**KEGG ORTHOLOGY** [Ortholog groups](#) [[KO system](#)]  
**KEGG GENOME** [Genomes](#) [[KEGG organisms](#)]  
**KEGG GENES** [Genes and proteins](#) [[Release history](#)]  
**KEGG COMPOUND** [Small molecules](#) [[Compound classification](#)]  
**KEGG REACTION** [Biochemical reactions](#) [[Reaction modules](#)]  
**KEGG DISEASE** [Human diseases](#) [[Cancer](#) | [Infectious disease](#)]  
**KEGG DRUG** [Drugs](#) [[ATC drug classification](#)]  
**KEGG MEDICUS** [Health information resource](#) [[Drug labels search](#)]

### ● Organism-specific entry points

**KEGG Organisms** Enter org code(s)   [hsa](#) [hsa eco](#)

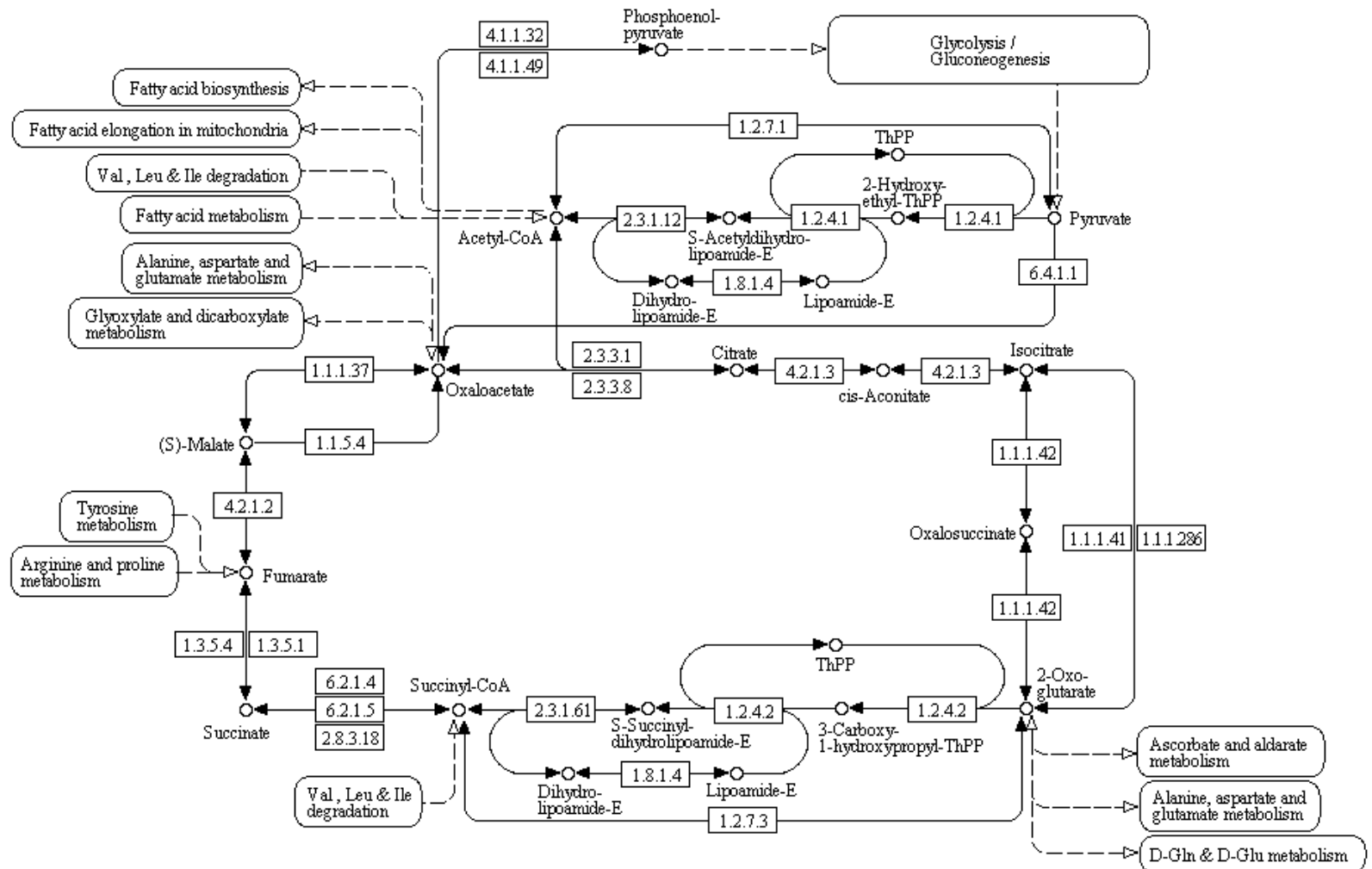
### ● Annotation and analysis tools

**KEGG Annotation** [KEGG Orthology \(KO\) based annotation](#) *New!*  
**BlastKOALA** [Automatic KO annotation service](#)  
**KEGG Mapper** [KEGG PATHWAY/BRITE/MODULE mapping tools](#)  
**KEGG Atlas** [Navigation tool to explore KEGG global maps](#)  
**BLAST/FASTA** [Sequence similarity search](#)  
**SIMCOMP** [Chemical structure similarity search](#)

# Exemple de voie métabolique

Voie de référence

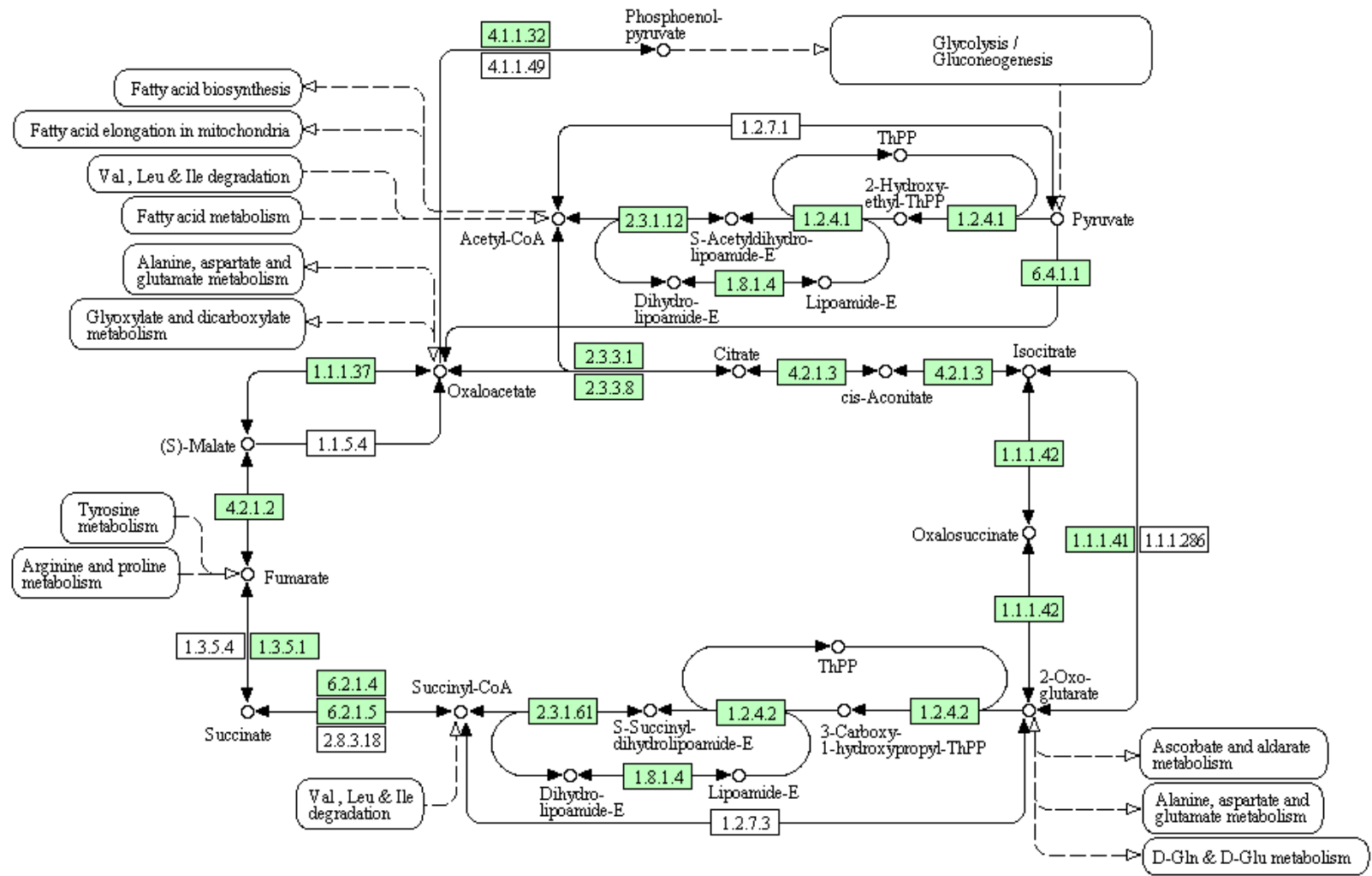
CITRATE CYCLE (TCA CYCLE)



# Exemple de voie métabolique

Voie chez l'homme

CITRATE CYCLE (TCA CYCLE)



# Nomenclature des enzymes

- Classification hiérarchique (4 niveaux) des activités enzymatiques
- Une réaction est référencée par un code : EC number

- 1 Oxidoreductases
- 2 Transferases
  - 2.1 Transferring one-carbon groups
    - 2.1.1 Methyltransferases
      - 2.1.1.1 nicotinamide *N*-methyltransferase
      - 2.1.1.2 guanidinoacetate *N*-methyltransferase
      - 2.1.1.3 thetin-homocysteine *S*-methyltransferase
      - 2.1.1.4 acetylserotonin *O*-methyltransferase
      - 2.1.1.5 betaine-homocysteine *S*-methyltransferase
    - ...
    - 2.8 Transferring sulfur-containing groups
    - 2.9 Transferring selenium-containing groups
- 3 Hydrolases
- 4 Lyases
- 5 Isomerases
- 6 Ligases

# Exemple de processus cellulaire : assemblage du flagelle



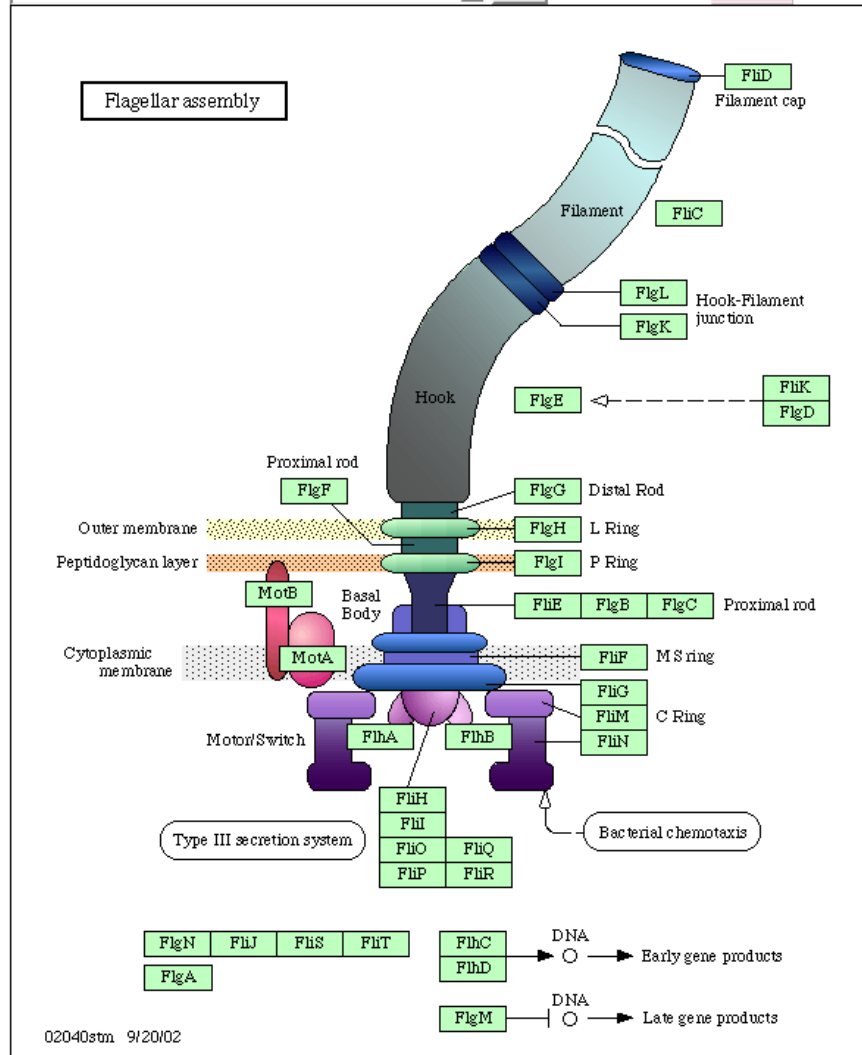
Flagellar assembly - Escherichia coli K-12 MG1655

[ [Pathway menu](#) | [Ortholog table](#) ]

Escherichia coli K-12 MG1655

Go

Current selection [Select](#)



# Autres ressources

PubMed : banque bibliographique (>30 millions de références)

OMIM : Online Mendelian Inheritance in Man : base de connaissances sur les maladies génétiques humaines

Gene Ontology : Vocabulaire structuré pour décrire les produits des gènes des différents organismes (PAS une banque mais un modèle des connaissances)



# Gene Ontology

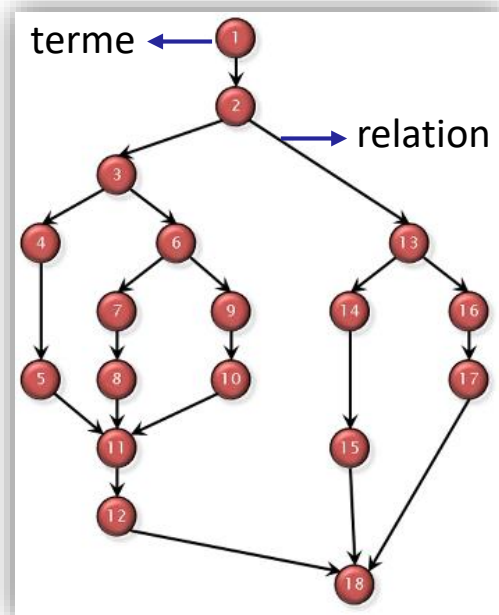
**Gene Ontology** est un projet destiné à structurer la description des gènes et des produits géniques en utilisant un vocabulaire contrôlé (un même terme pour décrire un même concept) et structuré commun à toutes les espèces. Cette structuration s'appelle une ontologie. Ce projet bio-informatique poursuit trois objectifs :

- gérer et enrichir le vocabulaire contrôlé décrivant les gènes et leurs produits,
- gérer les *annotations*, c'est-à-dire les informations rattachées aux gènes et à leurs produits,
- fournir les outils permettant d'accéder aux informations structurées dans le cadre du projet.

Gene ontology a été initialement créée en 1998 par un consortium de chercheurs étudiants le génome de trois organisme modèles : *Drosophila melanogaster*, *Mus musculus* et *Saccharomyces cerevisiae*. Depuis, d'autres bases de données sur des organismes modèles ont rejoint le consortium pour contribuer au développement du projet.

# Gene Ontology

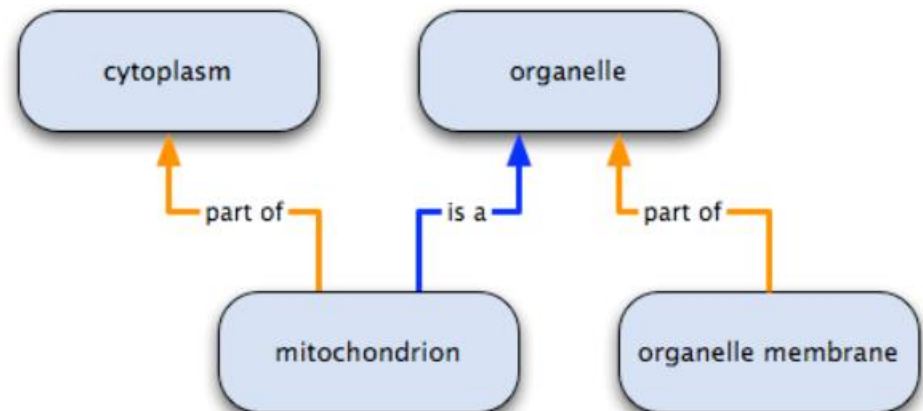
La base GO est conçue comme un graphe orienté acyclique, ce qui permet de représenter une hiérarchie. Chaque nœud du graphe représente un terme et les arêtes les relations entre les termes. Chaque terme est en relation avec un ou plusieurs termes du même domaine, et parfois d'autres domaines. Le vocabulaire GO est construit pour être indépendant des espèces considérées, avec des termes applicables à la fois aux organismes multicellulaires et unicellulaires, aux eucaryotes et aux procaryotes.



graphe acyclique orienté

Deux types de relation :

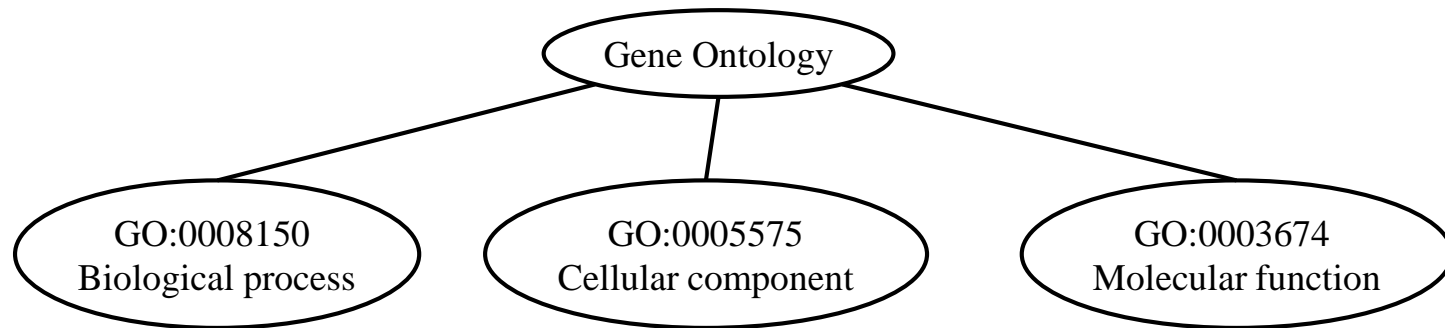
- *is a* : exemple la mitochondrie est un sous-type d'organelle (permet la spécification)
- *part of* : la membrane de l'organelle est une partie de l'organelle



# Gene Ontology

En fait trois ontologies :

- Cellular component
- Biological process
- Molecular function

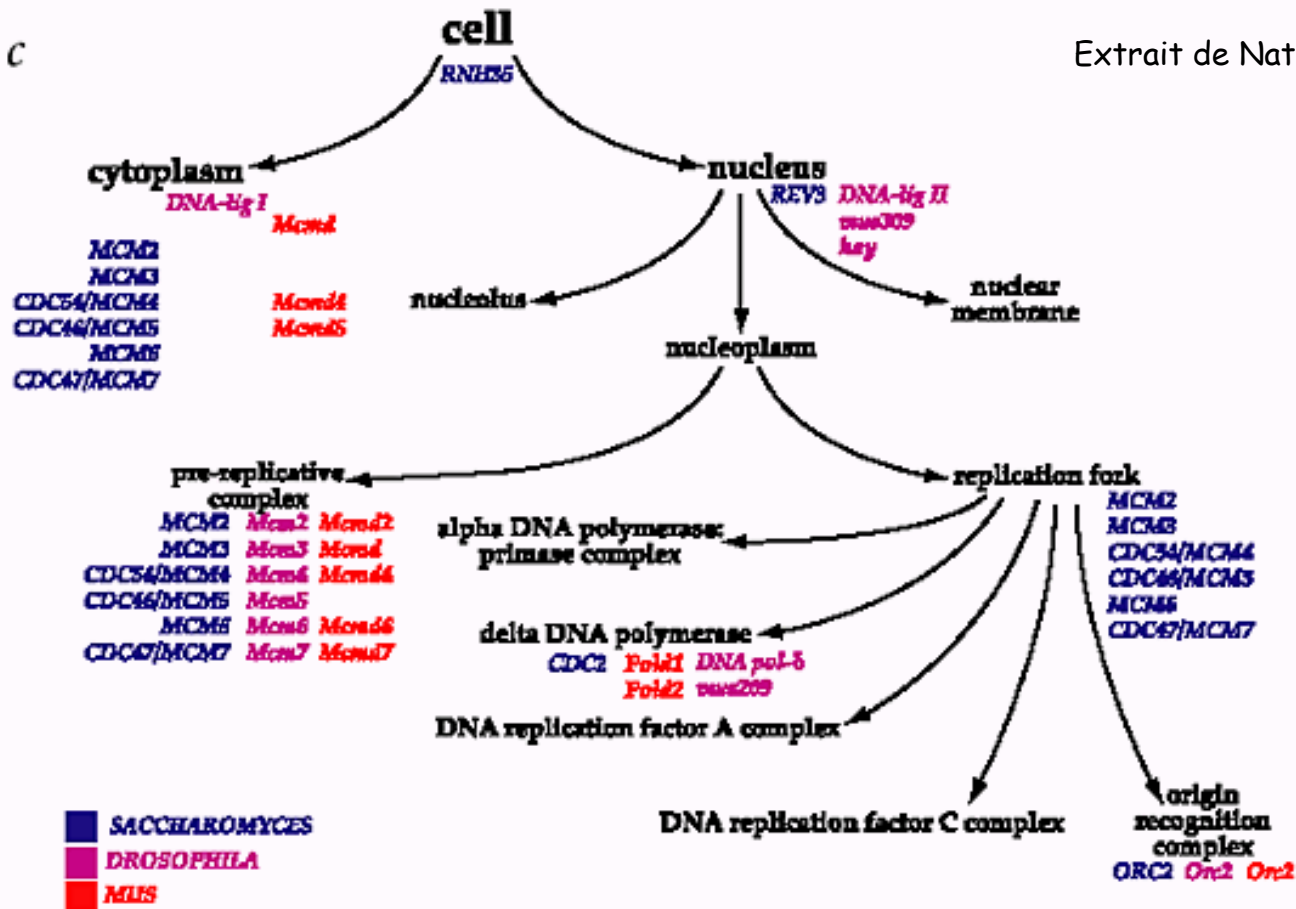


# Gene Ontology : Cellular component ontology

Ces termes décrivent un composant d'une cellule qui fait partie d'un objet plus grand, tel qu'une structure anatomique (par exemple le réticulum endoplasmique rugueux ou le noyau mais aussi un groupe de produit de gènes (par exemple le ribosome, le protéasome ou un dimère de protéine)

C

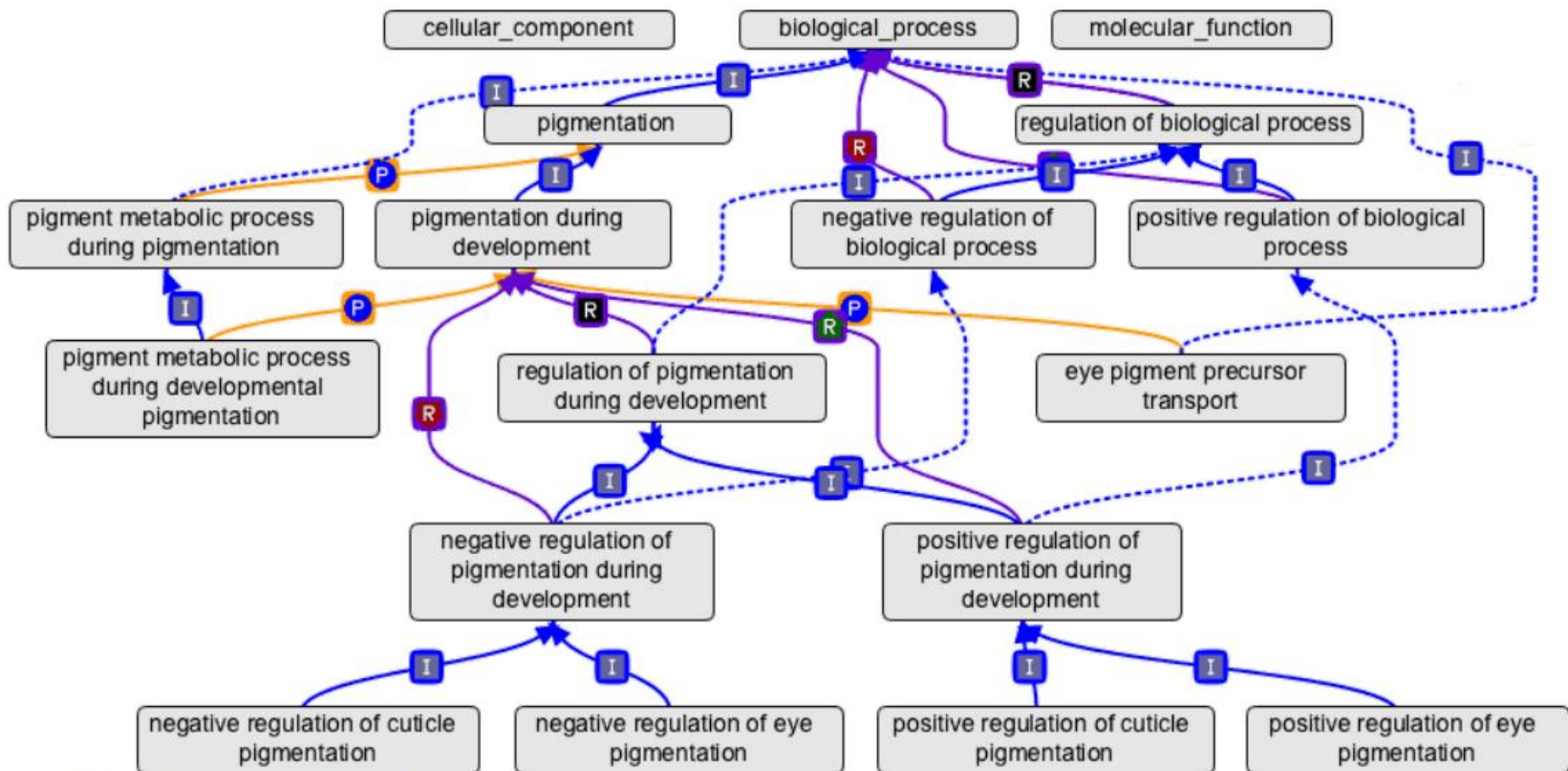
Extrait de Nature genetics (2000) 25, 25-29



# Gene Ontology : Biological process ontology

## Biological process ontology

Ces termes décrivent une série d'évènements accomplie par un ou plusieurs ensembles organisés de fonctions moléculaires. Pas équivalent à un « pathway » car n'essaie pas de représenter la dynamique ou les dépendances nécessaires pour décrire un « pathway ». Exemple de termes généraux « transduction de signal ». Exemple de termes plus spécifiques « processus métabolique de la pyrimidine ». Distinction entre processus biologique et fonction moléculaire : un processus doit avoir plusieurs étapes distinctes.



A

set of terms under the biological process node pigmentation.

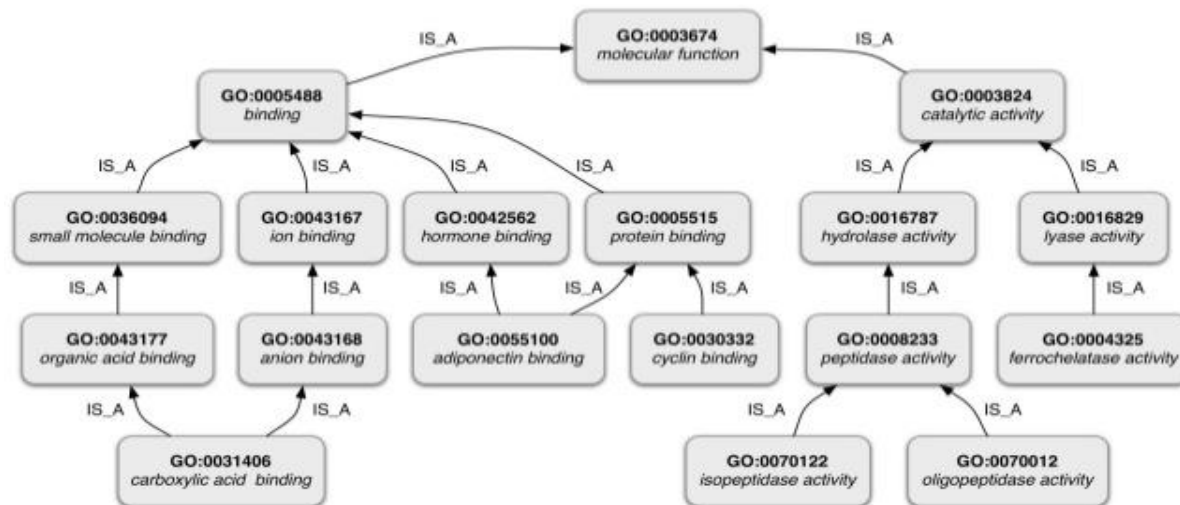
# Gene Ontology : Molecular function ontology

## Molecular function ontology

Ces termes décrivent des activités qui se produisent au niveau moléculaire, telles que "activité catalytique" ou "activité de liaison". Les fonctions moléculaires correspondent généralement à des activités pouvant être réalisées par des produits géniques individuels, mais certaines activités sont réalisées par des complexes. Exemples de termes fonctionnels généraux « activité catalytique » et « activité transporteur »; Exemples de termes fonctionnels plus étroits « activité adénylate cyclase » ou « la liaison au récepteur Toll ».

## Gene Ontology (GO)

Example from Molecular Function ontology



# Annotation Gene Ontology

- Un numéro unique représentant le terme
- L'ontologie (P pour Biological Process, C pour Cellular component et F pour Molecular Function)
- Une description

```
DR    GO; GO:0030175; C:filopodium; IDA.  
DR    GO; GO:0030027; C:lamellipodium; IDA.  
DR    GO; GO:0005886; C:plasma membrane; IDA.  
DR    GO; GO:0001726; C:ruffle; IDA.  
DR    GO; GO:0000902; P:cellular morphogenesis; IDA.  
DR    GO; GO:0030324; P:lung development; IMP.  
DR    GO; GO:0001946; P:lymphangiogenesis; IMP.  
DR    GO; GO:0051272; P:positive regulation of cell motility; IDA.
```

## Interrogation des bases de données

Les banques de données sont maintenues par divers instituts et organismes. Elles sont mises à disposition, le plus généralement, via un site Web (serveurs) proposant une interface de consultation et d'interrogation. De plus, ces serveurs proposent parfois des interfaces pour l'utilisation de logiciels en ligne.

Principaux sites :

- Serveur de l'EBI : European Bioinformatics Institute (Europe)
- Serveur du NCBI : National Center for Biotechnology Information (US)
  - **Entrez** (<http://www.ncbi.nlm.nih.gov/gquery/>)

Interrogation banques protéiques uniquement :

- Serveur Uniprot: <http://www.uniprot.org/>



## Search NCBI databases

[Help](#)

topoisomerase [keyword] AND streptococcus pneumoniae [organism]

Search

**Literature**

|                                |   |
|--------------------------------|---|
| <a href="#">Books</a>          | books and reports                               |
| <a href="#">MeSH</a>           | ontology used for PubMed indexing               |
| <a href="#">NLM Catalog</a>    | books, journals and more in the NLM Collections |
| <a href="#">PubMed</a>         | scientific & medical abstracts/citations        |
| <a href="#">PubMed Central</a> | full-text journal articles                      |

**Health**

|                               |  |
|-------------------------------|--|
| <a href="#">ClinVar</a>       | human variations of clinical significance        |
| <a href="#">dbGaP</a>         | genotype/phenotype interaction studies           |
| <a href="#">GTR</a>           | genetic testing registry                         |
| <a href="#">MedGen</a>        | medical genetics literature and links            |
| <a href="#">OMIM</a>          | online mendelian inheritance in man              |
| <a href="#">PubMed Health</a> | clinical effectiveness, disease and drug reports |

**Genomes**

|                             |   |
|-----------------------------|---|
| <a href="#">Assembly</a>    | genomic assembly information                      |
| <a href="#">BioProject</a>  | biological projects providing data to NCBI        |
| <a href="#">BioSample</a>   | descriptions of biological source materials       |
| <a href="#">Clone</a>       | genomic and cDNA clones                           |
| <a href="#">dbVar</a>       | genome structural variation studies               |
| <a href="#">Epigenomics</a> | epigenomic studies and display tools              |
| <a href="#">Genome</a>      | genome sequencing projects by organism            |
| <a href="#">GSS</a>         | genome survey sequences                           |
| <a href="#">Nucleotide</a>  | DNA and RNA sequences                             |
| <a href="#">Probe</a>       | sequence-based probes and primers                 |
| <a href="#">SNP</a>         | short genetic variations                          |
| <a href="#">SRA</a>         | high-throughput DNA and RNA sequence read archive |
| <a href="#">Taxonomy</a>    | taxonomic classification and nomenclature catalog |

**Genes**

|                              |  |
|------------------------------|--|
| <a href="#">EST</a>          | expressed sequence tag sequences                       |
| <a href="#">Gene</a>         | collected information about gene loci                  |
| <a href="#">GEO DataSets</a> | functional genomics studies                            |
| <a href="#">GEO Profiles</a> | gene expression and molecular abundance profiles       |
| <a href="#">HomoloGene</a>   | homologous gene sets for selected organisms            |
| <a href="#">PopSet</a>       | sequence sets from phylogenetic and population studies |
| <a href="#">UniGene</a>      | clusters of expressed transcripts                      |

**Proteins**

|                                   |   |
|-----------------------------------|---|
| <a href="#">Conserved Domains</a> | conserved protein domains                         |
| <a href="#">Protein</a>           | protein sequences                                 |
| <a href="#">Protein Clusters</a>  | sequence similarity-based protein clusters        |
| <a href="#">Structure</a>         | experimentally-determined biomolecular structures |

**Chemicals**

|                                   |  |
|-----------------------------------|--|
| <a href="#">BioSystems</a>        | molecular pathways with links to genes, proteins and chemicals |
| <a href="#">PubChem BioAssay</a>  | bioactivity screening studies                                  |
| <a href="#">PubChem Compound</a>  | chemical information with structures, information and links    |
| <a href="#">PubChem Substance</a> | deposited substance and chemical information                   |

# Résultat de la requête

## Search NCBI databases

[Help](#)

topoisomerase [keyword] AND streptococcus pneumoniae [organism] Search

About 608 search results for "topoisomerase [keyword] AND streptococcus pneumoniae [organism]"

### Literature

|                |     |   |
|----------------|-----|---|
| Books          | 1   | books and reports                               |
| MeSH           | 0   | ontology used for PubMed indexing               |
| NLM Catalog    | 0   | books, journals and more in the NLM Collections |
| PubMed         | 0   | scientific & medical abstracts/citations        |
| PubMed Central | 346 | full-text journal articles                      |

### Health

|               |   |  |
|---------------|---|--|
| ClinVar       | 0 | human variations of clinical significance        |
| dbGaP         | 0 | genotype/phenotype interaction studies           |
| GTR           | 0 | genetic testing registry                         |
| MedGen        | 0 | medical genetics literature and links            |
| OMIM          | 0 | online mendelian inheritance in man              |
| PubMed Health | 0 | clinical effectiveness, disease and drug reports |

### Genomes

|             |   |   |
|-------------|---|---|
| Assembly    | 0 | genomic assembly information                      |
| BioProject  | 0 | biological projects providing data to NCBI        |
| BioSample   | 0 | descriptions of biological source materials       |
| Clone       | 0 | genomic and cDNA clones                           |
| dbVar       | 0 | genome structural variation studies               |
| Epigenomics | 0 | epigenomic studies and display tools              |
| Genome      | 0 | genome sequencing projects by organism            |
| GSS         | 0 | genome survey sequences                           |
| Nucleotide  | 0 | DNA and RNA sequences                             |
| Probe       | 0 | sequence-based probes and primers                 |
| SNP         | 0 | short genetic variations                          |
| SRA         | 0 | high-throughput DNA and RNA sequence read archive |
| Taxonomy    | 0 | taxonomic classification and nomenclature catalog |

### Genes

|              |     |  |
|--------------|-----|--|
| EST          | 0   | expressed sequence tag sequences                       |
| Gene         | 220 | collected information about gene loci                  |
| GEO DataSets | 0   | functional genomics studies                            |
| GEO Profiles | 0   | gene expression and molecular abundance profiles       |
| HomoloGene   | 0   | homologous gene sets for selected organisms            |
| PopSet       | 0   | sequence sets from phylogenetic and population studies |
| UniGene      | 0   | clusters of expressed transcripts                      |

### Proteins

|                   |    |   |
|-------------------|----|---|
| Conserved Domains | 0  | conserved protein domains                         |
| Protein           | 6  | protein sequences                                 |
| Protein Clusters  | 12 | sequence similarity-based protein clusters        |
| Structure         | 23 | experimentally-determined biomolecular structures |

### Chemicals

|                   |   |  |
|-------------------|---|--|
| BioSystems        | 0 | molecular pathways with links to genes, proteins and chemicals |
| PubChem BioAssay  | 0 | bioactivity screening studies                                  |
| PubChem Compound  | 0 | chemical information with structures, information and links    |
| PubChem Substance | 0 | deposited substance and chemical information                   |

# Résultat de la requête dans la banque Protein

NCBI Resources How To Sign in to NCBI

Protein Protein topoisomerase [keyword] AND streptococcus pneumoniae [organism] Search

Save search Advanced Help

Show additional filters

- Species  
Bacteria (6)  
More ...
- Source databases  
UniProtKB / Swiss-Prot (6)  
More ...

Sequence length  
Custom range...

Molecular weight  
Custom range...

Release date  
Custom range...

Revision date  
Custom range...

Clear all

Show additional filters

Display Settings: Summary, 20 per page, Sorted by Default order

## Results: 6

- [RecName: Full=DNA gyrase subunit B \[Streptococcus pneumoniae R6\]](#)  
1. 648 aa protein  
Accession: P0A4M0.1 GI: 61225464  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- [RecName: Full=DNA gyrase subunit B \[Streptococcus pneumoniae TIGR4\]](#)  
2. 648 aa protein  
Accession: P0A4L9.1 GI: 61225463  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- [RecName: Full=DNA gyrase subunit A \[Streptococcus pneumoniae R6\]](#)  
3. 822 aa protein  
Accession: Q8DPM2.1 GI: 30913085  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- [RecName: Full=DNA gyrase subunit A \[Streptococcus pneumoniae TIGR4\]](#)  
4. 822 aa protein  
Accession: P72524.3 GI: 17377446  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- [RecName: Full=DNA topoisomerase 4 subunit A; AltName: Full=Topoisomerase IV subunit A \[Streptococcus pneumoniae TIGR4\]](#)  
5. 823 aa protein  
Accession: P72525.3 GI: 19861240  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)
- [RecName: Full=DNA topoisomerase 4 subunit B; AltName: Full=Topoisomerase IV subunit B \[Streptococcus pneumoniae TIGR4\]](#)  
6. 647 aa protein  
Accession: Q59961.3 GI: 19864342  
[GenPept](#) [FASTA](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Filters: Manage Filters

## Results by taxon

- Top Organisms [Tree]
- Streptococcus pneumoniae TIGR4 (4)
  - Streptococcus pneumoniae R6 (2)

## Analyze these sequences

- Run BLAST
- Align sequences with COBALT
- Identify Conserved Domains with CD-Search

## Find related data

Database: Select

Find items

## Search details

topoisomerase [keyword] AND "Streptococcus pneumoniae" [Organism]

Search

See more...

## Recent activity

Turn Off Clear

- topoisomerase[keyword] AND streptococcus pneumoniae[orgar Protein]
- topoisomerase[All Fields] AND streptococcus pneumoniae[organi: Gene]
- topoisomerase[keyword] AND

# Récupération des séquences dans un fichier

Protein

Protein

topoisomerase [keyword] AND streptococcus pneumoniae [organism]

Search

Save search Advanced

Help

Show additional filters

Display Settings: Summary, 20 per page, Sorted by Default order

Send to: Filters: Manage Filters

Species

Bacteria (6)

More ...

Source

databases

UniProtKB / Swiss-Prot (6)

More ...

Sequence length

Custom range...

Molecular

weight

Custom range...

Release date

Custom range...

Revision date

Custom range...

Clear all

Show additional filters

## Results: 6

RecName: Full=DNA gyrase subunit B [Streptococcus pne

1. 648 aa protein

Accession: P0A4M0.1 GI: 61225464

GenPept FASTA Graphics Related Sequences Identical Pro

RecName: Full=DNA gyrase subunit B [Streptococcus pne

2. 648 aa protein

Accession: P0A4L9.1 GI: 61225463

GenPept FASTA Graphics Related Sequences Identical Pro

RecName: Full=DNA gyrase subunit A [Streptococcus pneumoniae R6]

3. 822 aa protein

Accession: Q8DPM2.1 GI: 30913085

GenPept FASTA Graphics Related Sequences Identical Proteins

RecName: Full=DNA gyrase subunit A [Streptococcus pneumoniae TIGR4]

4. 822 aa protein

Accession: P72524.3 GI: 17377446

GenPept FASTA Graphics Related Sequences Identical Proteins

RecName: Full=DNA topoisomerase 4 subunit A; AltName: Full=Topoisomerase IV subunit A [Streptococcus pneumoniae TIGR4]

5. 823 aa protein

Accession: P72525.3 GI: 19861240

### Choose Destination

- File  Clipboard  
 Collections  Analysis Tool

Download 6 items.

Format

FASTA

Sort by

Default order

Create File

### Find related data

Database: Select

Find items

### Search details

topoisomerase[keyword] AND "Streptococcus pneumoniae" [Organism]

# Séquences récupérées

## Format Fasta

```
>gi|61225464|sp|P0A4M0.1|GYRB_STRR6 RecName: Full=DNA gyrase subunit B
MTEEIKNLQAQDYDASQIQVLEGLEAVRMRPGMYIGSTSKEGLHHLVWEIVDNSIDEALAGFASHIQVFI
EPDDSITVVDDGRGIPVDIQEKTGRPAVETVFTVLHAGGKFGGGGYKVS GGLHGVGSSVVNALSTQLDVH
VHKNNGKIHYQEYRRGHVVADLEIVGDTDKTGTTVHFPTDPKIFTETTIFDFDKLNKRIQELAFNRLGLQI
SITDKRQGLEQTKHYHYEGGIASVVEYINENKDVI FDTPIYTDGEMDDITVEVAMQYTTGYHENVMSFAN
NIHTHEGGTHEQGFR TALTRVINDYARKNKLLKDNEDNLTGEDVREGLTAVISVKHPNPQFEGQTKTKLG
NSEVVKITNRLFSEAFSDFLMENPQIAKRIVEKGILAAKARVAAKRAREVTRKKSGLAISNLPGLADCS
SNNPAETELFIVEGDSAGGSAKSGRNREFQAILPIRGKILNVEKASMDKILANEEIRSLFTAMGTGFGAE
FDVSKARYQKLVMLTDADVDGAHIRTLTLLLTLIYRYMKPILEAGYVYIAQPPIYGKVGSEIKEYIQPGAD
QEIKLQEALARYSEGR TKPTIQRYKGLGEMDDHQLWETTMDPEHRLMARVSVDDAAEADKIFDMLMGDRV
EPRREFIEENAVYSTLDV
```

```
>gi|61225463|sp|P0A4L9.1|GYRB_STRPN RecName: Full=DNA gyrase subunit B
MTEEIKNLQAQDYDASQIQVLEGLEAVRMRPGMYIGSTSKEGLHHLVWEIVDNSIDEALAGFASHIQVFI
EPDDSITVVDDGRGIPVDIQEKTGRPAVETVFTVLHAGGKFGGGGYKVS GGLHGVGSSVVNALSTQLDVH
VHKNNGKIHYQEYRRGHVVADLEIVGDTDKTGTTVHFPTDPKIFTETTIFDFDKLNKRIQELAFNRLGLQI
SITDKRQGLEQTKHYHYEGGIASVVEYINENKDVI FDTPIYTDGEMDDITVEVAMQYTTGYHENVMSFAN
NIHTHEGGTHEQGFR TALTRVINDYARKNKLLKDNEDNLTGEDVREGLTAVISVKHPNPQFEGQTKTKLG
NSEVVKITNRLFSEAFSDFLMENPQIAKRIVEKGILAAKARVAAKRAREVTRKKSGLAISNLPGLADCS
SNNPAETELFIVEGDSAGGSAKSGRNREFQAILPIRGKILNVEKASMDKILANEEIRSLFTAMGTGFGAE
FDVSKARYQKLVMLTDADVDGAHIRTLTLLLTLIYRYMKPILEAGYVYIAQPPIYGKVGSEIKEYIQPGAD
QEIKLQEALARYSEGR TKPTIQRYKGLGEMDDHQLWETTMDPEHRLMARVSVDDAAEADKIFDMLMGDRV
EPRREFIEENAVYSTLDV
```

```
>gi|30913085|sp|Q8DPM2.1|GYRA_STRR6 RecName: Full=DNA gyrase subunit A
MQDKNLVNVNLTKEMKASFIDYAMSVIVARALPDVRDGLKPVHRRILYGMNELGVTPDKPHKKSARITGD
VMGKYHPHGDSSIYEAMVRMAQWWSYRYMLVDGHGNGFSMDGDSAAAQRYTEARMSKIALEMLRDINKNT
VDFVDNYDANEREPLVLPARFPNLLVNGATGIAVGMATNIPPHNLGETIDAVKLVMDNPEVTTKDLMEVL
PGPDFPTGALVMGKSGIHKAYETGKGSIVLRSRTEIETTKTGRERIVVTEFPYMVNKTKVHEHIVRLVQE
KRIEGITAVRDESNREGVRFVIEVKRDASANVILNNLKFMTQMQTNFGFNMLAIQNGIPKILSLRQILDA
YIEHQKEVVVRRTRFDKEKAEARAHILEGLLIALDHIDEVIRIRASETDAEAQAE LMSKFKLSERQSQ
ILDMRLRRLTG LERDKIQSEYDDLALIALDLADILAKPERVSQIIKDELDEVKRFSDKRRTELMVGQVL
SLEDEDLIEESDVLITLSNRGYIKRLDQDEFTAQKRGGRGVQGTGVKDDDFVRELVSTSTHDHLLFFTNK
GRVYRLKGYEIP EYGR TAKGLPVVNLK LDEDESIQTVINVESDRSDDAYLFFTTTRHGIVKRTSVKEFAN
IRQNGLKALNLKDEDELINVLLAEGDMDIIIGTKFGYAVRFNQS AVRGMSRIATGVKGVNLRREGD TVVGA
SLITDQDEVLIITEKGYGKRTVATEYPTKGRGGKGMQTAKITEKNGLLAGLMTVQGDEDLMIITDTGVMI
RTNLANISQTGRATMGVKVMRLDQDAQIVTFTTVAVAEKEEVGTENETEGEA
```

```
>gi|17377446|sp|P72524.3|GYRA_STRPN RecName: Full=DNA gyrase subunit A
MQDKNLVNVNLTKEMKASFIDYAMSVIVARALPDVRDGLKPVHRRILYGMNELGVTPDKPHKKSARITGD
```



# Interrogation sur le site d'UniProtKB : choix de la base de données

The image shows a screenshot of the UniProt website. At the top left is the UniProt logo. Below it are navigation links: BLAST, Align, Retrieve/ID. The main heading reads: "The mission of UniProt is to provide a central resource for protein sequence and functional information." On the left side, there is a section titled "UniProtKB" with two sub-sections: "Swiss-Prot (547,357)" described as "Manually annotated and reviewed" and "TrEMBL (89,451,166)" described as "Automatically annotated and not reviewed." The top right features a search bar with "UniProtKB" selected in a dropdown menu, an "Advanced" search option, and a search icon. A dropdown menu is open, showing three main categories: "UniProtKB" (Protein knowledgebase), "UniRef" (Sequence clusters), and "UniParc" (Sequence archive). Below these are three columns of links: "Proteomes" (Protein sets from fully sequenced genomes), "Supporting data" (Literature citations, Taxonomy, Keywords, Subcellular locations, Cross-referenced databases, Diseases, Annotation programs), and "Help" (Help pages, FAQs, UniProtKB manual, documents, news archive, etc.). At the bottom right, there are social media icons for Twitter, Facebook, and RSS, and a "News archive" link.

UniProtKB

BLAST Align Retrieve/ID

The mission of UniProt is to provide a central resource for protein sequence and functional information.

UniProtKB

Swiss-Prot (547,357)  
Manually annotated and reviewed.

TrEMBL (89,451,166)  
Automatically annotated and not reviewed.

UniProtKB

Protein knowledgebase

UniRef

Sequence clusters

UniParc

Sequence archive

Help Contact

Proteomes

Protein sets from fully sequenced genomes

Supporting data

Select one of the options below to target your search:

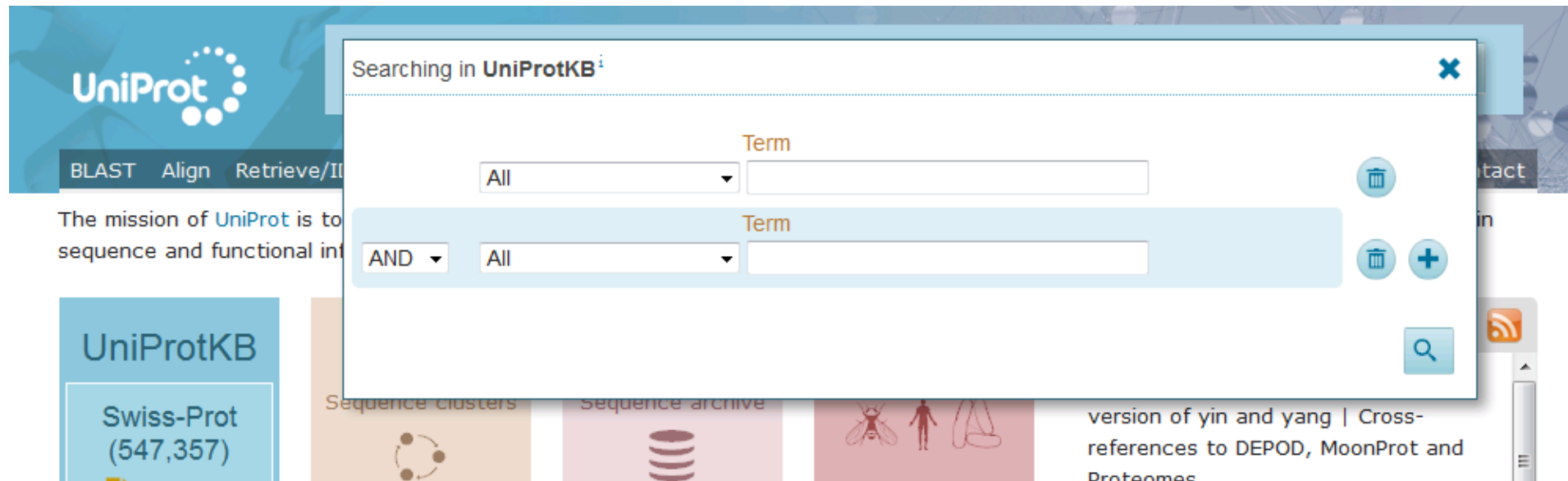
- Literature citations
- Taxonomy
- Keywords
- Subcellular locations
- Cross-referenced databases
- Diseases
- Annotation programs

Help

Help pages, FAQs, UniProtKB manual, documents, news archive, etc.

News archive

# Interrogation sur le site d'UniProtKB : construction de la requête



The image shows a screenshot of the UniProtKB search interface. The main header features the UniProt logo and navigation links for BLAST, Align, and Retrieve/Integrate. Below the header, the mission statement of UniProt is visible. The search interface is highlighted, showing a search box with a dropdown menu set to 'All' and a search button. The search box is titled 'Searching in UniProtKB' and has a close button (X) in the top right corner. The search box contains the text 'All' and a search button. Below the search box, there are several sections: 'UniProtKB' with a sub-section for 'Swiss-Prot (547,357)', 'Sequence clusters', 'Sequence archive', and a section for 'version of yin and yang | Cross-references to DEPOD, MoonProt and Proteomes'.

Dans le menu déroulant, choisir le champ de la fiche dans lequel on veut rechercher le terme.

All signifie que la recherche du terme se fera dans toute la fiche

# Interrogation sur le site d'UniProtKB : construction de la requête

The screenshot displays the UniProtKB search interface. At the top left, the UniProt logo is visible. Below it, there are navigation links for 'BLAST', 'Align', and 'Retrieve/II'. A brief description states: 'The mission of UniProt is to sequence and functional int...'. On the left side, a box highlights 'UniProtKB' with 'Swiss-Prot (547,357)' and 'Manually annotated and reviewed.' Below this is the 'TrEMBL' section. The main search area is titled 'Searching in UniProtKB'. It features a search bar with the text 'Streptococcus pneumoniae' and a dropdown menu set to 'Organism [OS]'. Below this, a search criteria box shows 'AND' as the operator, 'Family and Domains' as the category, and 'Protein family' as the sub-category. The search term 'topoisomerase' is entered in the search bar. A search button with a magnifying glass icon is located at the bottom right of the search area. To the right of the search area, there are icons for a trash can and a plus sign. Below the search area, a purple box labeled 'Supporting data' contains links for 'Literature citations', 'Taxonomy', and 'Subcellular locations'. On the right side of the page, there is a section for 'Proteomes' with a link to 'UniProt release 2015\_01'. At the bottom right, there are links for 'Higher and higher | New mouse and zebrafish variation files | Structuring'.

La recherche de *Streptococcus pneumoniae* restreint au champ Organism  
La recherche du terme topoisomerase dans le champ Protein Family



# Interrogation sur le site d'UniProtKB : resultat de la requête

UniProtKB  Advanced

BLAST Align Retrieve/ID Mapping Help Contact

Show help for UniProtKB

## Results

◦ Add columns: Protein families

Columns BLAST Align Download Add to basket 1 to 25 of 1,384 Show 25

| Entry                           | Entry name | Protein names   | Gene names            | Organism  | Length |
|---------------------------------|------------|---|-----------------------|---|--------|
| <input type="checkbox"/> Q59961 | PARE_STRPN | <b>DNA topoisomerase 4 subunit B</b> (EC 5.99.1.3) (Topoisomerase IV subunit B) | <b>parE</b> , SP_0852 | Streptococcus pneumoniae serotype 4 (strain ATCC BAA-334 / TIGR4) | 647    |
| <input type="checkbox"/> P72525 | PARC_STRPN | <b>DNA topoisomerase 4 subunit A</b> (EC 5.99.1.3) (Topoisomerase IV subunit A) | <b>parC</b> , SP_0855 | Streptococcus pneumoniae serotype 4 (strain ATCC BAA-334 / TIGR4) | 823    |
| <input type="checkbox"/> P0A4L9 | GYRB_STRPN | <b>DNA gyrase subunit B</b> (EC 5.99.1.3)                                       | <b>gyrB</b> , SP_0806 | Streptococcus pneumoniae serotype 4 (strain ATCC BAA-334 / TIGR4) | 648    |
| <input type="checkbox"/> P0A4M0 | GYRB_STRR6 | <b>DNA gyrase subunit B</b> (EC 5.99.1.3)                                       | <b>gyrB</b> , spr0715 | Streptococcus pneumoniae (strain ATCC BAA-255 / R6)               | 648    |

**Filter by**

- Reviewed (6) Swiss-Prot
- Unreviewed (1,378) TrEMBL

**Popular organisms**

- STRPN (5)
- STRR6 (5)
- Streptococcus pneumoniae SPN034156 (5)
- Streptococcus pneumoniae SPN034183 (5)
- Streptococcus pneumoniae SPN994038 (5)

Other organisms

**Search terms**

Filter

Distribution des séquences en fonction de la banque de données

1384 séquences versus 6 sur le site du NCBI

6 dans la banque SwissProt expertisée

# Récupération des séquences dans un fichier



UniProtKB

organism:"streptococcus pneumoniae" family:topoisomerase ANE

Advanced



BLAST Align Retrieve/ID Mapping

Help Contact

Show help for UniProtKB

## Results

o Add columns: Protein families

Basket

### Filter by<sup>i</sup>

Reviewed (6)  
Swiss-Prot

### Popular organisms

STRPN (4)

STRR6 (2)

### Search terms

Filter  
"topoisomerase" as:  
protein family

### View by

Columns BLAST Align Download Add to basket

1 to 6 of 6

Show 25

| Entry  | Entry name | Format                        | Length |
|--------|------------|-------------------------------|--------|
| Q59961 | PARE_STRPN | FASTA (canonical)             | 647    |
| P72525 | PARC_STRPN | DNA topoisomerase 4 subunit A | 823    |

Download selected (0)

Download all (6)

Format: FASTA (canonical) Go

Preview first 10

# Séquences récupérées

## Format Fasta

```
>sp|Q59961|PARE_STRPN DNA topoisomerase 4 subunit B OS=Streptococcus pneumoniae serotype 4 |(
MSKKEININNYNDDAIQVLEGLDAVRKRPGMYIGSTDGAGLHHLVWEIVDNAVDEALSGF
GDRIDVTINKDGLTVQDHGRGMPTGMHAMGIPTVEVIFTILHAGGKFGQGGYKTSGLLH
GVGSSVVNALSSWLEVEITRDGAVYKQRFENGKPVTTLKKIGTAPKSKTGKVTFMPDA
TIFSTDFKYNTISERLINESAFLLKNVTLSTLTKRTNEAIEFHYENGVQDFVSYLNEDKE
ILTPVLYFEGEDNGFQVEVALQYNDGFSNLSLFSVNNVRTKDGGETHETGLKSAITKVMND
YARKTGLLKEKDKNLEGS DYREGLA AVL SILVPEEHLQFEGQTKDKLGSPLARPVVDGIV
ADKLTFFLMENGELASNLIRKAIKARDAREAAARKARDES RNKGKKNKKDKGLLSGKLT PAQ
SKNPAKNELYLVEGDSAGGSAKQGRDRKFQAILPLRGKVNTAKAKMADILKNEEINTMI
YTIGAGVGADFSIEDANYDKIIIMTDADTDGAHIQTLTLLTFFYRYMRPLVEAGHVYIALP
PLYKMSKGGKKEEVAYAWTDGELEELRQFGKGATLQRYKGLGEMNADQLWETTMNPET
RTLIRVTIEDLARAERRVNVLMGDKVEPRRkWIEDNVKFTLEETTF
>sp|P72525|PARC_STRPN DNA topoisomerase 4 subunit A OS=Streptococcus pneumoniae serotype 4 (
MSNIQNMSLEDIMGERFGRYSKYIIQDRALPDIRDGLKPVQRRILYSMNKDSNTFDKSYR
KSAKSVGNIMGNFHPHGDSSIIDAMVRMSQNWKNREILVEMHGNGSMDGDPFAAMRYTE
ARLSEIAGYLLQDIEKKTVPFAWNFDDTEKEPTVLPAAFPNLLVNGSTGISAGYATDIPP
HNLAEVIDA AVY MIDHPTAKIDKLMFELPGDPFPTGAI IQGRDEIKKAYETGKGRVVVRS
KTEIEKLGKKEQIVII EIPYEINKANLVKIDDV RVNKNVAGIAEVRDES DRDGLRIAI
ELKKDANTELVNLVLFKYTDLQINYNFMVAIDNFTPRQVGI VPI LSSYIAHREVI LAR
SRFDKEKA EKRLHIVEGLRVISILDEVIALIRASENKADAKENLKVSYDFTEEQAEAI V
TLQLYRLTNTD VVVLQEEEAELREKIAMLAAIIGDERTMYNLMKKELEREVKKKFATPRLS
SLEDTAKAIEIDTASLIAEEDTYVSVTKAGYIKRTSPRSFAASTLEEIGKRDDDRLIFVQ
SAKTTQHLLMFTSLGNVIYRPIHELADIRWKDIGEHLSTITNFETNEEILYVEVLDQFD
DATTYFAVTRLGQIKRVERKEFTPWRTYRSKSVKYAKLKDDTDQIVAVAPIKLDVVVLS
QNGYALRFNIEEVPVVGAKAAGVKAMNLKEDDVLQSGFICNTSSFYLLTQRGSLKRVSI E
EILATSRAKRGLQVLRLELKNKPHRVFLAGAVAEQGFVGDFFSTEV DVNDQTL LVQSNKGT
IYESRLQDLNLSERTSNGSFISDTISDEEVFDAYLQEVVTEDEK
>sp|P0A4L9|GYRB_STRPN DNA gyrase subunit B OS=Streptococcus pneumoniae serotype 4 (strain AT
MTEEIKNLQAQDYDASQIQVLEGLEAVRMRPGMYIGSTSKEGLHHLVWEIVDNSIDEALA
GFASHIQVFIEPDDSI TVVDDGRGIPVDIQEKTGRPAVETVFTVLHAGGKFGGGYKVSG
GLHGVGSSVVNALSTQLDVHVHKNKGIHYQEYRRGHVVADLEIVGDTDKTGTTVHFTPDP
KIFTETTFDFDKLNKRIQELAFNLNRLQISITDKRQGLEQTKHYHYEGGIASIVEYINE
NKDVI FDTPIYTDGEMDDITVEVAMQYTTGYHENVMSFANNIHTHEGGTHEQGFR TALTR
VINDYARKNKLLKDNEDNLTGEDVREGLTAVISVKHPNPQFEGQTKTKLGNSEVVKITNR
LFSEAFSDFLMENPQIAKRIVEKGILAAKARVAARKRAREVTRKKSGL EISNLPGKLADCS
SNNPAETELFIVEGDSAGGSAKSGRNREFQAILPIRGKILNVEKASMDKILANEEIRSLF
TAMGTGFGAEFDVSKARYQKLVLMTDADVDGAHIRTLLLTLIYRYMKPILEAGVYVYIAQP
PIYGVKVGSEIKEYIQPGADQEIKLQEALARYSEGRTKPTIQRYKGLGEMDDHQLWETTM
DPEHRLMARVSVDDAAEADKIFDMLMGDRVEPRREFIENAVYSTLDV
>sp|P0A4M0|GYRB_STRR6 DNA gyrase subunit B OS=Streptococcus pneumoniae (strain ATCC BAA-255
MTEEIKNLQAQDYDASQIQVLEGLEAVRMRPGMYIGSTSKEGLHHLVWEIVDNSIDEALA
GFAHIQVFI
```

# Interconnexion des banques de données : références croisées

Définition : Lien (référence) d'une entrée d'une banque vers une entrée d'une (autre) banque.

```
LOCUS       CAA65843                  268 aa          linear   PLN 19-DEC-1997
DEFINITION  CBEL protein, formerly GP34 [Phytophthora parasitica].
ACCESSION   CAA65843
VERSION     CAA65843.1  GI:2706495
DBSOURCE    embl accession X97205.1
KEYWORDS    .
SOURCE      Phytophthora parasitica
  ORGANISM  Phytophthora parasitica
REFERENCE   1
  AUTHORS   Mateos,F.V., Rickauer,
  TITLE     Cloning and characteri
            Phytophthora parasitic
            cellulose-binding and
            JOURNAL Mol. Plant Microbe Int
            PUBMED 9390419
REFERENCE   2
  TITLE     Protein
            sig peptide
            Region
            1..268
            /product="CBEL protein, formerly GP34"
            /function="elicits defense reactions in host"
            1..18
            23..55
            /region_name="CBM 1"
            /note="Fungal cellulose binding domain; c102521"
            /db_xref="CDD:154956"
CDS         1..268
            /gene="cbel1"
            /coded_by="X97205.1:53..859"
            /note="localized in cell wall;
lectin-like and cellulose-binding properties"
            /db_xref="GOA:042830"
            /db_xref="InterPro:IPR000177"
            /db_xref="InterPro:IPR000254"
            /db_xref="InterPro:IPR003014"
            /db_xref="InterPro:IPR003609"
            /db_xref="UniProtKB/TrEMBL:042830"
ORIGIN
```