Banques et bases de données

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

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 (>100 milliards de bp, >100 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'accession**

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RL   Unpublished.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
```

Exemple d'entrée dans la banque EMBL (suite)

```
FH   Key             Location/Qualifiers
FH
FT   source          1..48787
FT                   /organism="Mus musculus"
FT                   /strain="129/SV"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:10090"
FT   promoter        1..9640
FT                   /gene="E11"
FT   mRNA            join(9641..9915,33318..33454,35464..35620,39063..39101,
FT                   40324..40435,41199..42193)
FT                   /gene="E11"
FT                   /product="transmembrane glycoprotein E11"
FT   exon            9641..9915
FT                   /gene="E11"
FT                   /number=1
FT   CDS             join(9849..9915,33318..33454,35464..35620,39063..39101,
FT                   40324..40435,41199..41205)
FT                   /codon_start=1
FT                   /gene="E11"
FT                   /product="transmembrane glycoprotein E11"
FT                   /note="gp38; PA2.26; OTS8"
FT                   /db_xref="GOA:Q62011"
FT                   /db_xref="InterPro:IPR008783"
FT                   /db_xref="MGI:103098"
FT                   /db_xref="UniProtKB/Swiss-Prot:Q62011"
FT                   /protein_id="AAM66761.1"
FT                   /translation="MWTVPVLFWVLGSVWFWDSAQGGTIGVNEDDIVTPGTGDGMVPPG
FT                   IEDKITTTGATGGLNESTGKAPLVPTQRERGTKPPLEELSTSATSDHDHREHESTTTVK
FT                   VVTSHSVDKKTSHPNRDNAGDETQTTDKKDGLPVVTLVGIIVGVLLAIGFVGGIFIVVM
FT                   KKISGRFSP"
FT   exon            33318..33454
FT                   /gene="E11"
FT                   /number=2
FT   exon            35464..35620
FT                   /gene="E11"
FT                   /number=3
FT   exon            39063..39101
FT                   /gene="E11"
FT                   /number=4
FT   exon            40324..40435
FT                   /gene="E11"
FT                   /number=5
FT   exon            41199..42193
FT                   /gene="E11"
FT                   /number=6
FT   polyA_signal    42172..42177
FT                   /gene="E11"
```

Exemple d'entrée dans la banque EMBL

```
SQ   Sequence 48787 BP; 12689 A; 11292 C; 11677 G; 13094 T; 35 other;
     tatagtcact cttgccaatt gcactaagga taacccaacc ttggtttaaa aaaaaaaatc        60
     ctgactcaaa gatgaatttc acacttagca gagacattta cctgccaaga 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 tacacgtggt 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
     caaaggtccc acctctgaat gacacaaccc cgtcgcttat 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 ggtttgtctt ctgatttcca tgcgcatgag cacggatatg cacacatata     48660
     cctgcataca cacacacaca cacacacaca cacacacaca agcatactcg tgcacatatg     48720
     tgcattcata taaacacata cacacaaaaa tgaaccttat cttatttaat tactttttt      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
```

Exemple d'entrée dans la banque GenBank (suite)

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

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: > 4 milliards d'a.a., > 12 millions de séquences

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;
     MWTVPVLFWV LGSVWFWDSA QGGTIGVNED DIVTPGTGDG MVPPGIEDKI TTTGATGGLN
     ESTGKAPLVP TQRERGTKPP 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

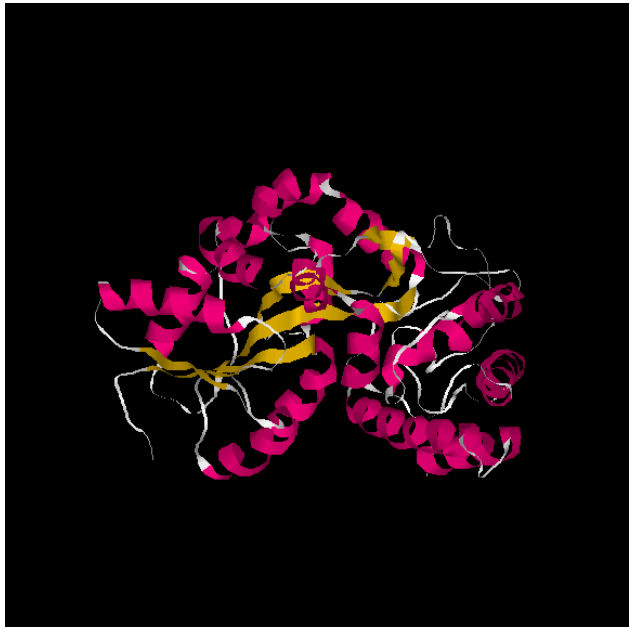
Une entrée contient donc les coordonnées de tous les atomes de la structure

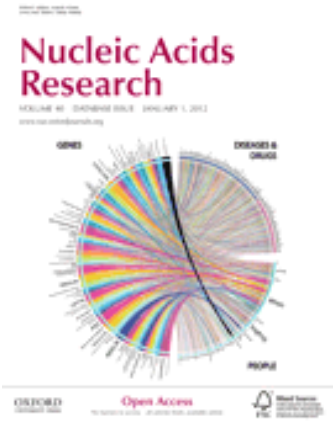
Exemple d'entrée dans la banque PDB

```
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 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection

Michael Y. Galperin and Xosé M. Fernández-Suárez

The 19th annual Database Issue of *Nucleic Acids Research* features descriptions of **92 new online databases** covering various areas of molecular biology and 100 papers describing recent updates to the databases previously described in *NAR* and other journals. The highlights of this issue include, among others, a description of neXtProt, a knowledgebase on human proteins; a detailed explanation of the principles behind the NCBI Taxonomy Database; NCBI and EBI papers on the recently launched BioSample databases that store sample information for a variety of database resources; descriptions of the recent developments in the Gene Ontology and UniProt Gene Ontology Annotation projects; updates on Pfam, SMART and InterPro domain databases; update papers on KEGG and TAIR, two universally acclaimed databases that face an uncertain future; and a separate section with 10 wiki-based databases, introduced in an accompanying editorial. The *NAR* online Molecular Biology Database Collection, available at http://www.oxfordjournals.org/nar/database/a/, has been updated and now lists **1380 databases**. Brief machine-readable descriptions of the databases featured in this issue, according to the BioDBcore standards, will be provided at the http://biosharing.org/biodbcore web site. The full content of the Database Issue is freely available online on the *Nucleic Acids Research* web site (http://nar.oxfordjournals.org/).

Banques/bases de données spécialisées

Quelques exemples :

Dédiées à un organisme :

- Flybase : Drosophile
- HIV database : virus HIV
- Colibri : *Escherichia coli*
- Subtilis : *Bacillus subtilis*
- plusieurs pour *Arabidopsis thaliana* et pour le riz
- etc

Dédiées à un type de séquences particulier :

- IMGT : données d'immunologie
- EPD : Eukaryotic Promoter Database
- ISFINDER : séquences d'insertion dans les bactéries et les archaebactéries
- The European ribosomal RNA database
- ABCdb : ABC transporteurs de bactéries et d'archaebactéries
- Transportdb : transporteurs
- etc

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
  - …

Go Clear

» Japanese

KEGG Home
Introduction
Overview
Release notes
Current statistics

KEGG Identifiers
Pathway maps
Brite hierarchies

KEGG XML

KEGG API

KEGG FTP

KegTools

GenomeNet

DBGET/LinkDB

Feedback

New features
Module maps
Disease/drug maps
3rd global map
Combined maps

KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, the ecosystem, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents	Update notes	Help

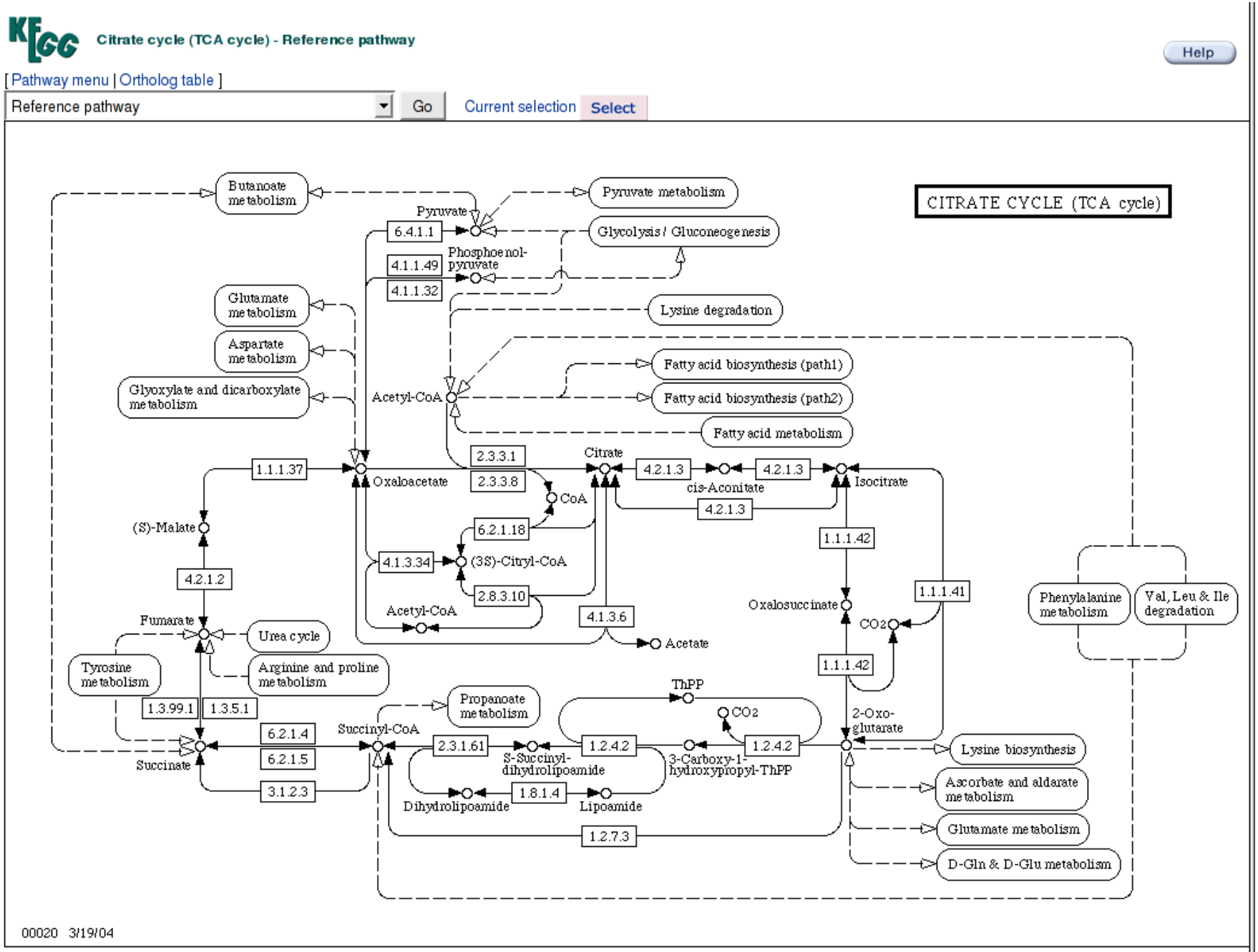
Data-oriented entry points

KEGG PATHWAY	Pathway maps for systemic functions	Pathway maps
KEGG BRITE	Functional hierarchies and ontologies	Brite hierarchies
KEGG MODULE	Module maps for functional units	KEGG modules
KEGG DISEASE	Human diseases	Disease classification
KEGG DRUG	Drugs	ATC drug classification
KEGG ORTHOLOGY	KO system and ortholog annotation	KO system
KEGG GENES	Genes and proteins	
KEGG GENOME	Genomes	KEGG organisms
KEGG COMPOUND	Chemical compounds	Compound classification
KEGG GLYCAN	Glycans	
KEGG REACTION	Reactions	

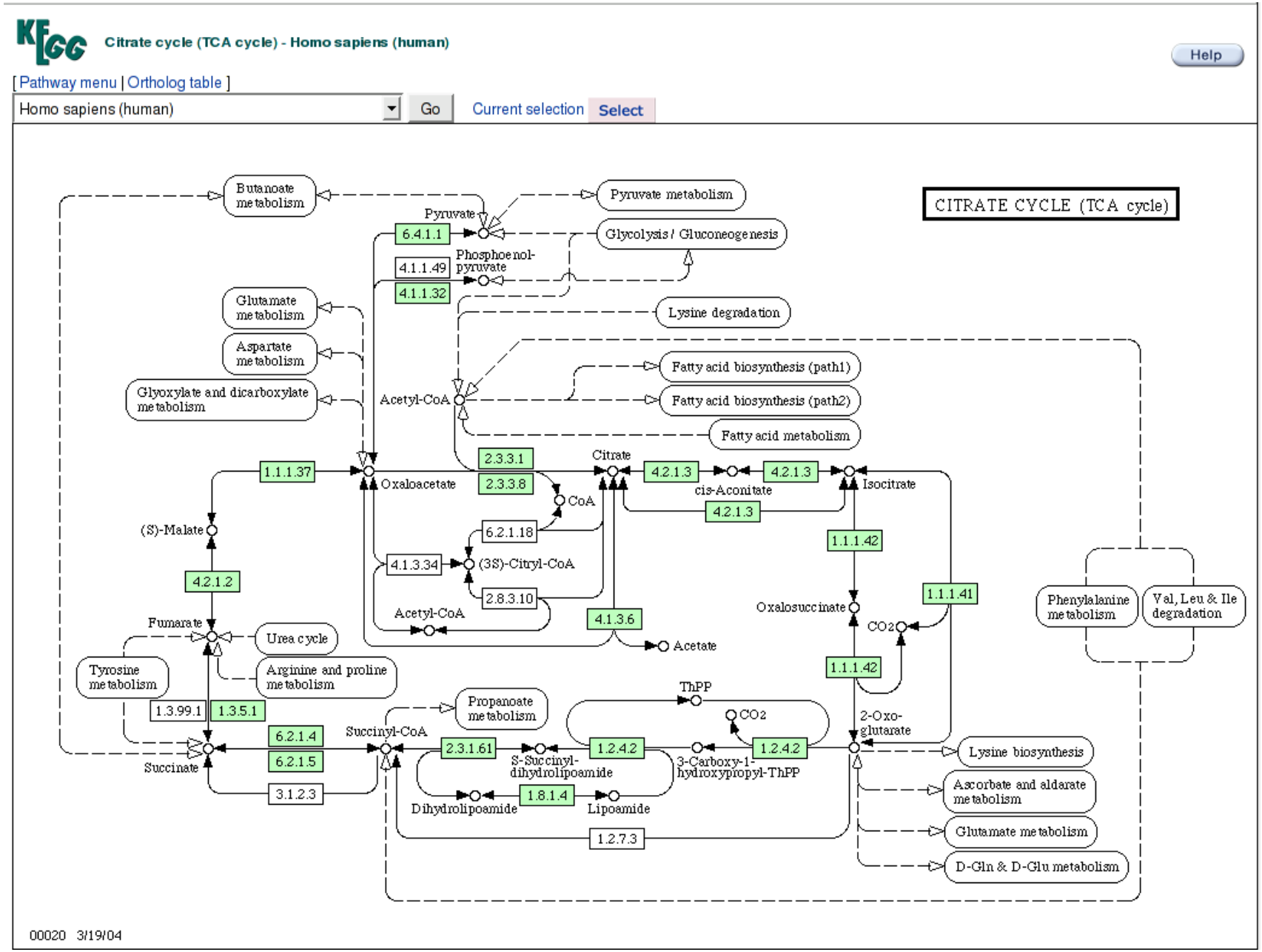
Organism-specific entry points

KEGG Organisms Select Organism Go (example) hsa

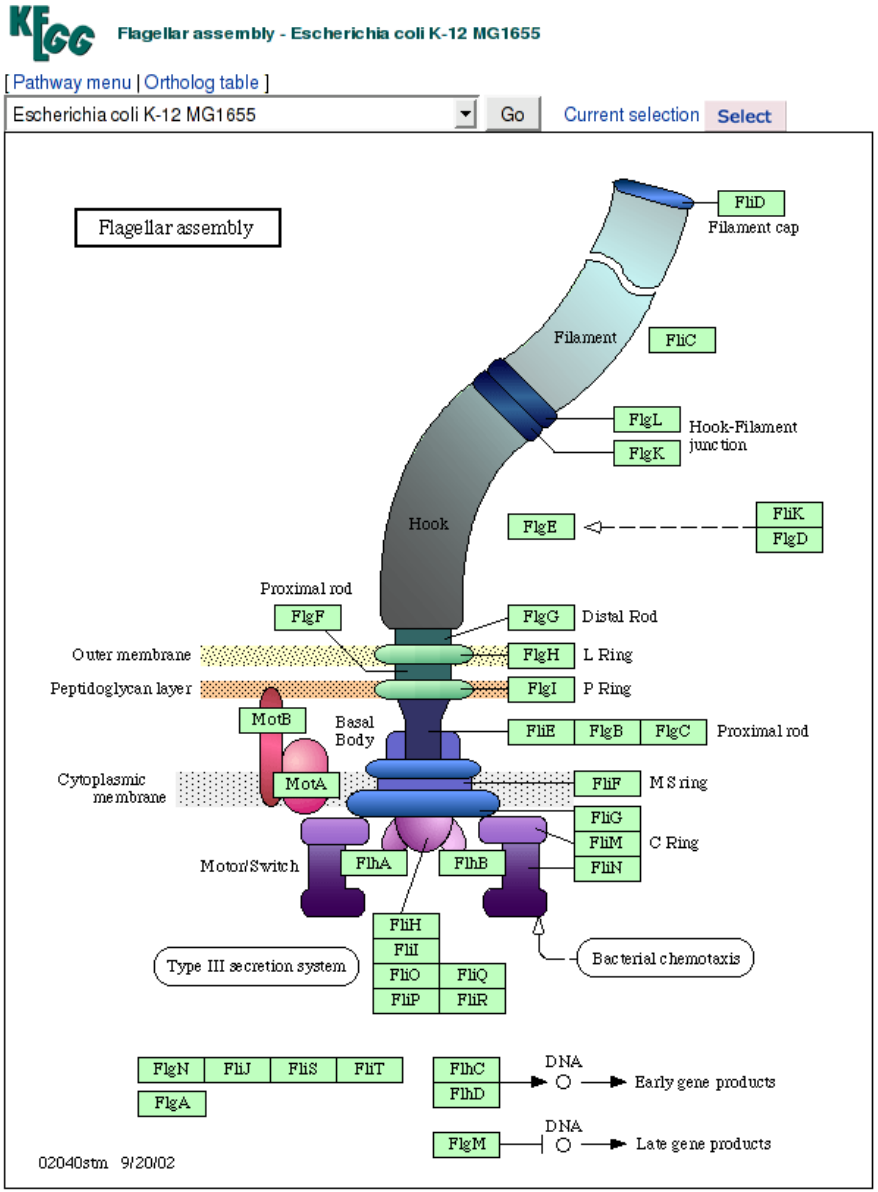
Exemple de voie métabolique



Exemple de voie métabolique



Exemple de processus cellulaire



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

Autres ressources

PubMed : banque bibliographique (>20 millions de références)

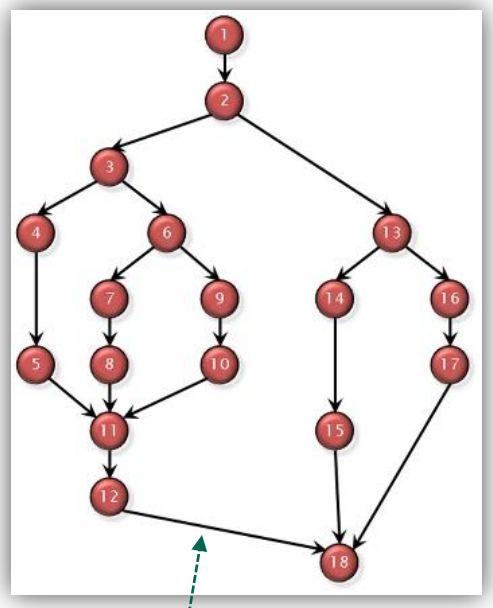
OMIM : Online Mendelian Inheritance in Man : base de connaissances sur les maladies génétiques humaines

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

Gene Ontology

ontologie : ensembles de termes reliés par des relations

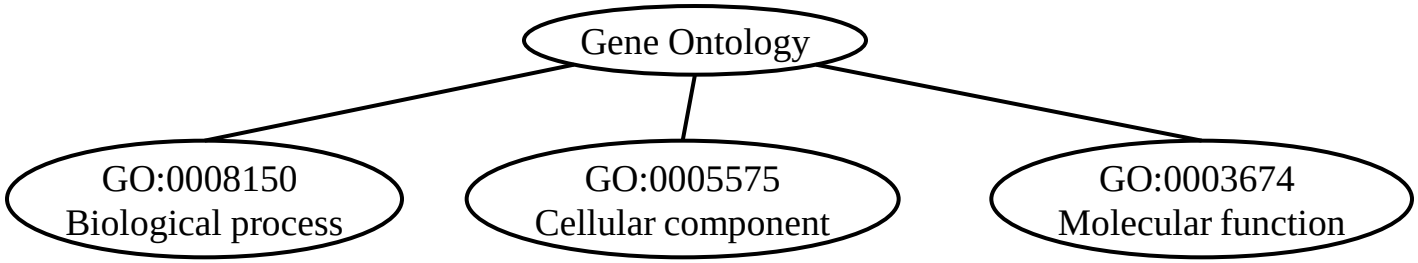
graphe acyclique orienté



relation de type :
is_a (est un)
ou
part_of (fait parti de)

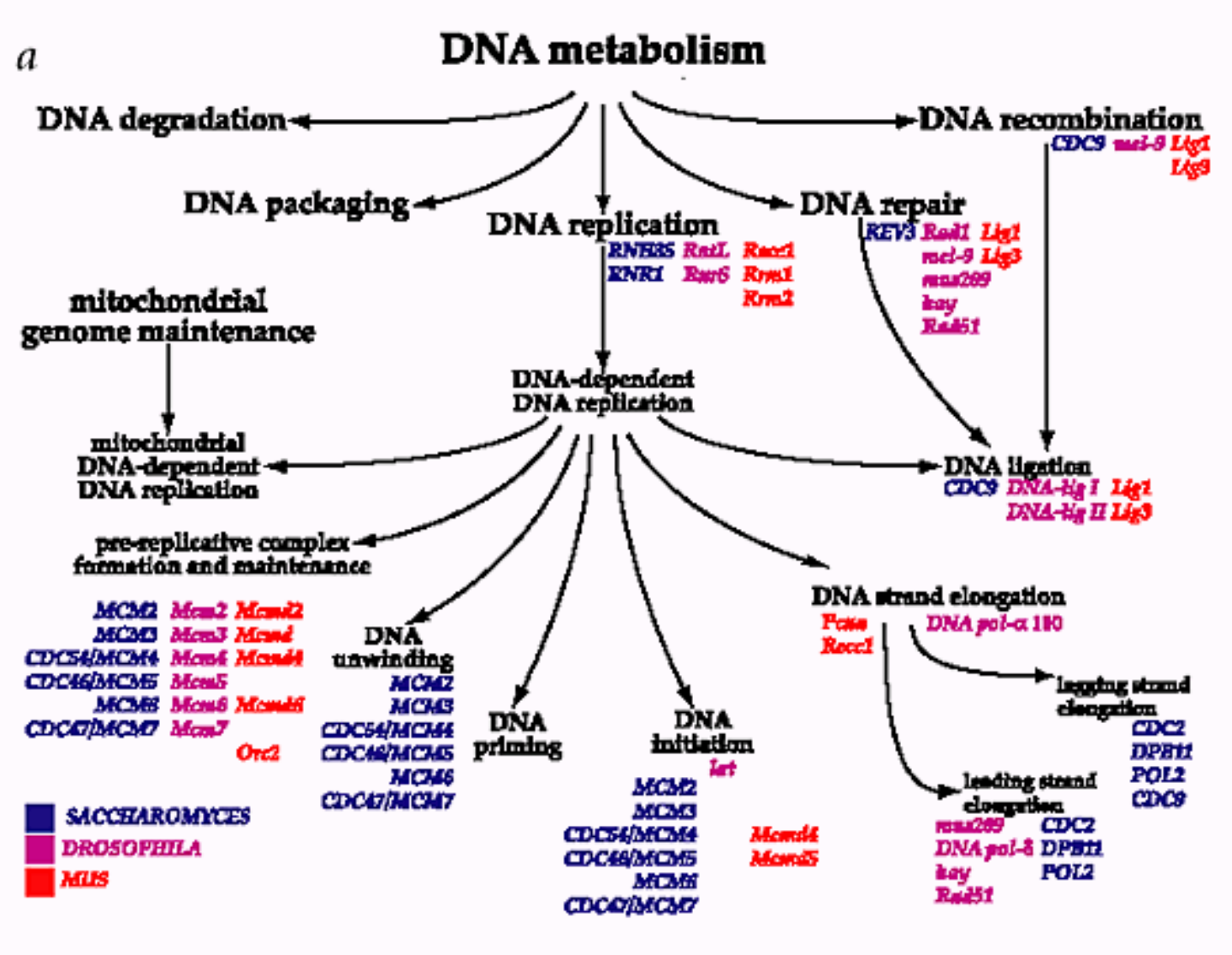
Gene Ontology

- Vocabulaire contrôlé (~20 000 termes)
  - Le même terme pour désigner le même concept
- Annotation du produit des gènes
- Structure hiérarchique : différents niveaux de détails
- 3 branches :
  - processus biologiques
  - fonctions moléculaires
  - composants cellulaires



Gene Ontology : Biological process ontology

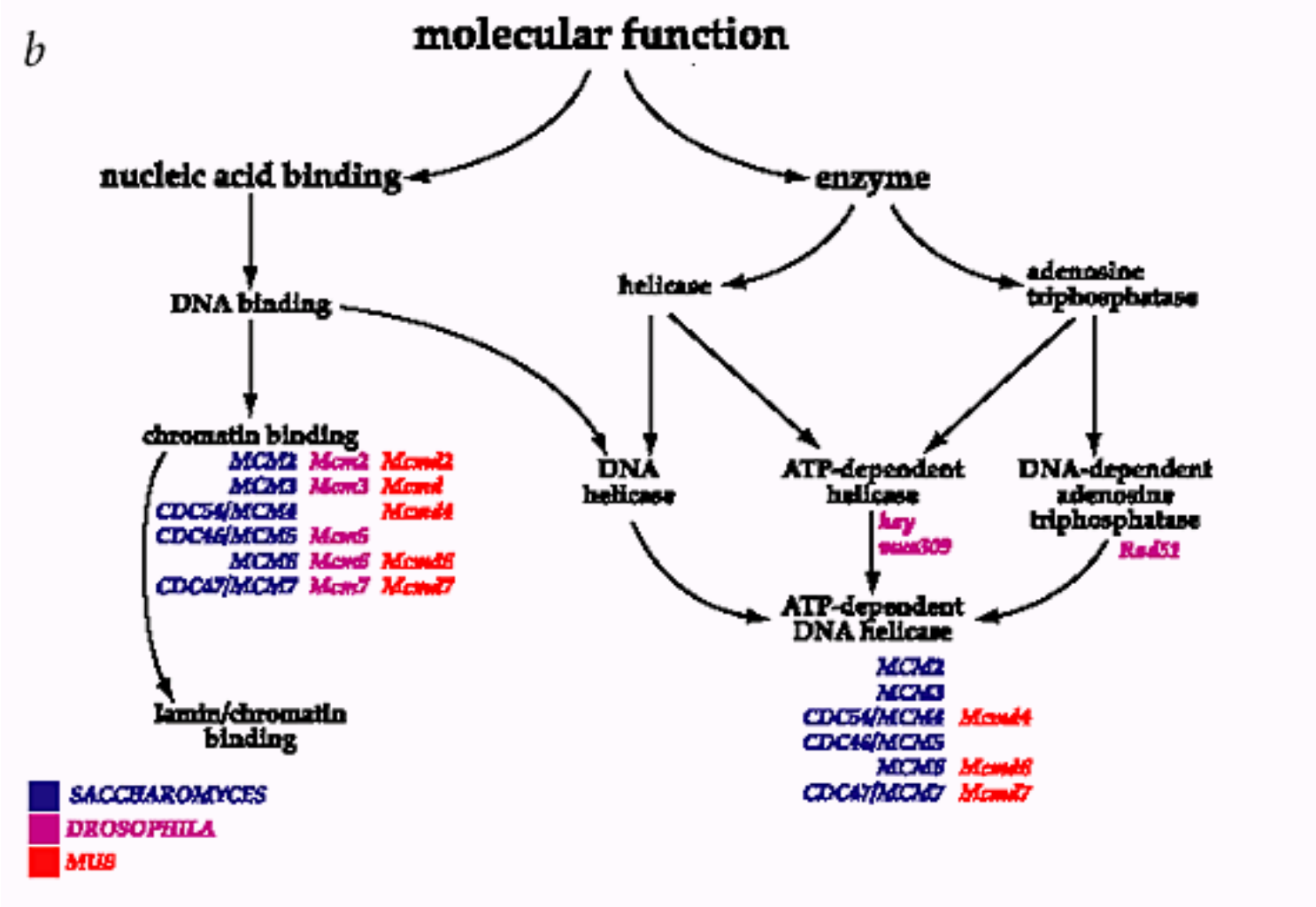
Une portion de l'ontologie décrivant le métabolisme de l'ADN



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

Gene Ontology : Molecular function ontology

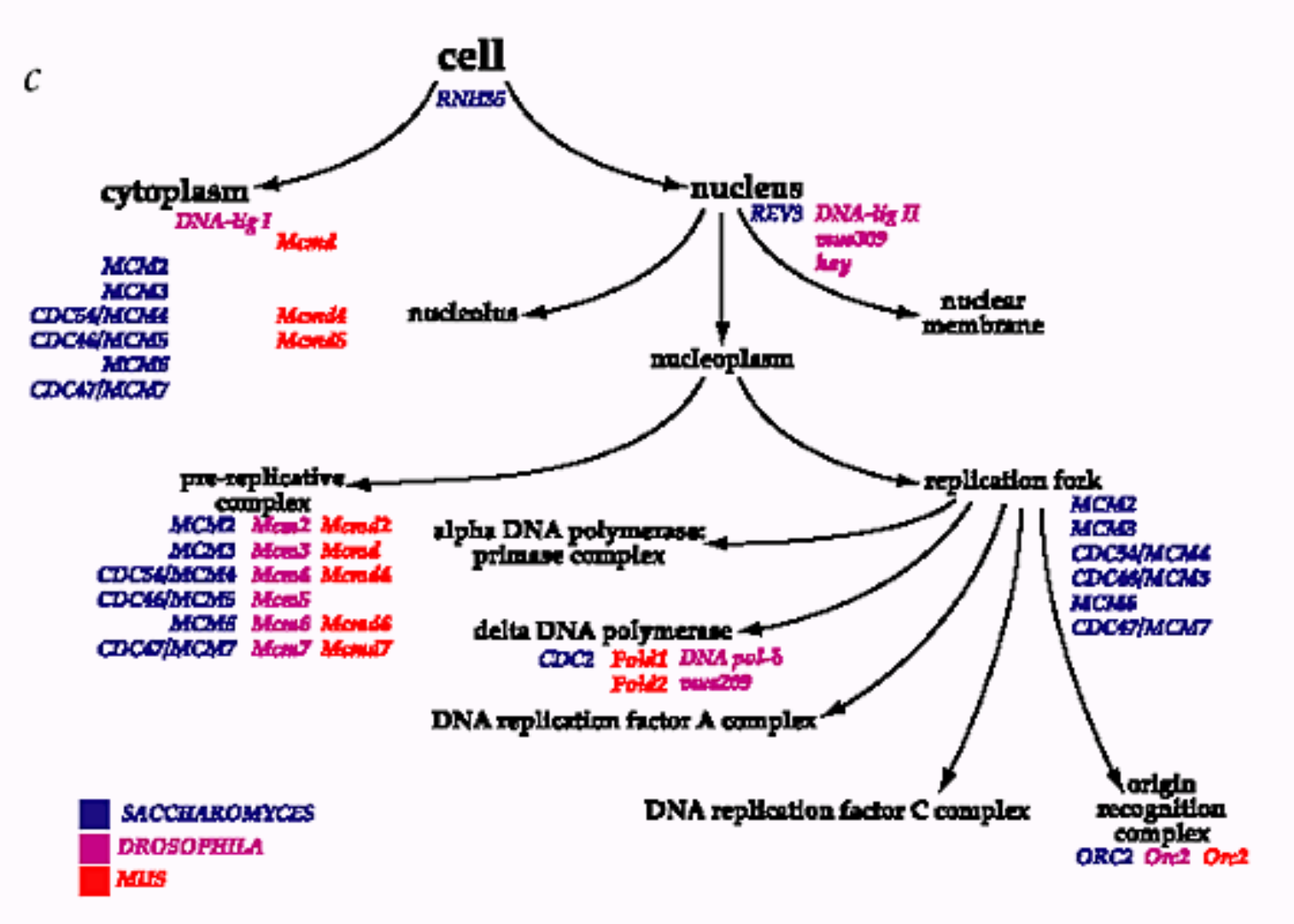
L'ontologie n'a pas été établie pour représenter des pathways réels mais pour refléter des catégories conceptuelles de la fonction des produits des gènes



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

Gene Ontology : Cellular component ontology

L'ontologie est définie de manière à représenter une cellule eucaryote générique.



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

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)
    - Interface d'interrogation : le système **SRS** (Sequence Retrieval System)

- Serveur du NCBI : National Center for Biotechnology Information (US)
    - **Entrez**

- Serveur ExPASy : Expert Protein Analysis System (Suisse)

SRS

EMBL-EBI
European Bioinformatics Institute

Quick Search | Library Page | Query Form | Tools | Results | Projects | Views | Databanks | HELP

Reset | Quick Search

Search Options

1. **Select** the **databanks** you want to search
2. Enter your **search terms** in the **Quick Search** box, or choose a **query form** from below

Standard Query Form
Extended Query Form

You can **browse** through all the **entries** in any **databanks**. First, **select** the **databanks** you want to browse, then click:

Browse Entries

Tips

- bookmark this **link** to return to your project
- *Linking to SRS?*

- Please read our Linking to SRS guide for important information regarding linking to our SRS server.

BookMarkLets

About BookmarkLets

- Protein Seq
- DNA/RNA Seq
- Structures

Available Databanks

Expand all | Collapse all | Show databanks tooltips: ☑

- Literature, Bibliography and Reference Databases
  - all ☐ TAXONOMY ☐ GENETICCODE ☐ OMIM ☐ MEDLINE
  - ☐ Patent Abstracts ☐ Karyn's Genomes
  - ***Literature, Bibliography and Reference Databases - subsections***
  - all ☐ MEDLINE (Updates) ☐ MEDLINE (Main Release 2006) ☐ MED2PUB
- Gene Dictionaries and Ontologies
- Nucleotide sequence databases
  - all ☐ EMBL ☐ Patent DNA ☐ IMGT/LIGM-DB ☐ EMBL (Coding Sequences)
  - ☐ IMGT/HLA ☐ IPD-KIR ☐ EMBL (Contig) ☐ Genome Reviews
  - ☐ EMBL (Contigs expanded) ☐ EMBL (Annotated Cons) ☐ RefSeq Genome ☐ LiveLists
  - ☐ EMBL ID/Accession Mapping ☐ EMBL MGA
  - ***Nucleotide sequence databases - subsections***
  - all ☐ EMBL (Updates) ☐ EMBL (Release) ☐ EMBL (Whole Genome Shotgun)
  - ☐ EMBL (Whole Genome Shotgun release) ☐ EMBL (Whole Genome Shotgun updates) ☐ EMBL (Contig release)
  - ☐ EMBL (Contig updates) ☐ EMBL (Contigs expanded release) ☐ EMBL (Contigs expanded updates)
  - ☐ EMBL (Annotated Cons release) ☐ EMBL (Annotated Cons updates) ☐ RefSeq Genome (Release)
  - ☐ RefSeq Genome (Updates) ☐ EMBL (Whole Genome Shotgun Masters)
- Nucleotide related databases
- UniProt Universal Protein Resource
  - all ☐ UniProtKB ☑ UniProtKB/Swiss-Prot ☐ UniProtKB/TrEMBL ☐ UniRef100 ☐ UniRef90
  - ☐ UniRef50 ☐ UniParc
- Other protein sequence databases
- Protein function, structure and interaction databases
- Enzymes, reactions and metabolic pathway databases
- Mutation and SNP databases
- Biological Resources Catalogues (CABRI)
- Mapping databases
- Other databases
- User owned databases
- Application result databases
- EMBOSS result databases
- Eurofir Food data
- EMBLCDS Grouped By

SRS

EMBL-EBI
European Bioinformatics Institute

Quick Search | Library Page | Query Form | Tools | Results | Projects | Views | Databanks | HELP

Reset search UniProtKB/Swiss-Prot

Search Options

Combine search terms with: & (AND)

Use wildcards ☑

Get results of type: Entry

Result Display Options

View results using: UniprotView

or

Create a view

Show 30 results per page

Tips

To do more advanced queries, use the **Extended Query** Form.

Fields you can search	Your search terms
Description	topoisomerase
Organism Name	Streptococcus pneumoniae
AllText	
AllText	

In a single field, you can separate multiple values by &, |, ! Search

Create a view

Select the fields you want displayed in your view and choose the format

Choose 1 or more fields:

ID
EntryName
AccessionNumber
Creation Date
Seq Mod Date
Annot Mod Date
Description

Display As: Table List

Sequence Format: swiss

Search

SRS Release 7.1.3.2 Copyright ©1997-2003 LION bioscience AG. All Rights Reserved. Terms of Use Feedback & Support

SRS

EMBL-EBI
European Bioinformatics Institute

Quick Search | Library Page | Query Form | Tools | Results | Projects | Views | Databanks | HELP

Reset

Query '([swissprot-Description:topoisomerase*] & (([swissprot-Organism:Streptococcus*] & [swissprot-Organism:pneumoniae*]) | [swissprot-Organism:Streptococcus pneumoniae*])) ' found 2 entries

Apply Options to:

- selected results only
- unselected results only

Result Options

Launch analysis tool: BlastP Launch

Show tools relevant to these results: Tools

Link to related information: Link

Save results: Save

Display Options

View results using: UniprotView

Sort results by: unsorted

- ascending
- descending

Show 30 results per page

Printer friendly view

Apply Display Options

UniProtKB/Swiss-Prot	Accession	UniSave	Description	GeneName	Species	Keywords	SeqLength
UniProtKB/Swiss-Prot:PARC_STRPN	P72525	P72525	DNA topoisomerase 4 subunit A (EC 5.99.1.-) (Topoisomerase IV subunit A).	PARC	STREPTOCOCCUS PNEUMONIAE.	Complete proteome DNA-binding Isomerase Topoisomerase	823
UniProtKB/Swiss-Prot:PARE_STRPN	Q59961	Q59961	DNA topoisomerase 4 subunit B (EC 5.99.1.-) (Topoisomerase IV subunit B).	PARE	STREPTOCOCCUS PNEUMONIAE.	ATP-binding Complete proteome Isomerase Nucleotide-binding Topoisomerase	647

Entrez

NCBI

Entrez, The Life Sciences Search Engine

HOME | SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer | BLAST

Search across databases topoisomerase [keyword] AND streptococcus pneun GO Clear Help

- Result counts displayed in gray indicate one or more terms not found

none	**PubMed:** biomedical literature citations and abstracts	none	**Books:** online books
275	**PubMed Central:** free, full text journal articles	none	**OMIM:** online Mendelian Inheritance in Man
none	**Site Search:** NCBI web and FTP sites	none	**OMIA:** online Mendelian Inheritance in Animals

none	**Nucleotide:** Core subset of nucleotide sequence records	none	**dbGaP:** genotype and phenotype
none	**EST:** Expressed Sequence Tag records	none	**UniGene:** gene-oriented clusters of transcript sequences
none	**GSS:** Genome Survey Sequence records	none	**CDD:** conserved protein domain database
6	**Protein:** sequence database	60	**3D Domains:** domains from Entrez Structure
none	**Genome:** whole genome sequences	none	**UniSTS:** markers and mapping data
7	**Structure:** three-dimensional macromolecular structures	none	**PopSet:** population study data sets
none	**Taxonomy:** organisms in GenBank	none	**GEO Profiles:** expression and molecular abundance profiles
none	**SNP:** single nucleotide polymorphism	none	**GEO DataSets:** experimental sets of GEO data
none	**dbVar:** Genomic structural variation	none	**Epigenomics:** Epigenetic maps and data sets
103	**Gene:** gene-centered information	none	**Cancer Chromosomes:** cytogenetic databases
none	**SRA:** Sequence Read Archive	none	**PubChem BioAssay:** bioactivity screens of chemical substances

Entrez

NCBI Resources How To My NCBI Sign In

Protein
Translations of Life

Search: Protein Save search Limits Advanced search Help

topoisomerase [keyword] AND streptococcus pneumoniae [Organism] Search Clear

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

Gene Information

This search in Gene shows 103 results , including:

topA (*Streptococcus pneumoniae TCH8431/19A*): DNA topoisomerase I
SP_1263 (*Streptococcus pneumoniae TIGR4*): DNA topoisomerase I
topA (*Streptococcus pneumoniae D39*): DNA topoisomerase I

Gene

Results: 6

1. RecName: Full=DNA gyrase subunit B
   648 aa protein
   P0A4L9.1 GI:61225463
   GenPept FASTA Graphics Related Sequences Identical Proteins
2. RecName: Full=DNA topoisomerase 4 subunit B; AltName: Full=Topoisomerase IV subunit B
   647 aa protein
   Q59961.3 GI:19864342
   GenPept FASTA Graphics Related Sequences Identical Proteins
3. RecName: Full=DNA topoisomerase 4 subunit A; AltName: Full=Topoisomerase IV subunit A
   823 aa protein
   P72525.3 GI:19861240
   GenPept FASTA Graphics Related Sequences Identical Proteins
4. RecName: Full=DNA gyrase subunit A
   822 aa protein
   P72524.3 GI:17377446
   GenPept FASTA Graphics Related Sequences Identical Proteins

Filter your results:

All (6)
Bacteria (6)
Related Structures (6)
RefSeq (0)

Manage Filters

Top Organisms [Tree]
Streptococcus pneumoniae (6)
Streptococcus pneumoniae R6 (2)

Analyze these sequences
Run BLAST
Align sequences with COBALT

Find related data
Database: Select
Find items

Search details

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

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 parasitic
            Eukaryota; stramenopil
REFERENCE   1
  AUTHORS   Mateos,F.V., Rickauer,
  TITLE     Cloning and characteri
            Phytophthora parasitic
            cellulose-binding and
  JOURNAL   Mol. Plant Microbe Int
   PUBMED   9390419
```

```
FEATURES             Location/Qualifiers
     source          1..268
                     /organism="Phytophthora parasitica"
                     /cultivar="nicotianae"
                     /db_xref="taxon:4792"
                     /clone="B 311.2"
                     /tissue_type="mycelium"
                     /clone_lib="lambda Zap II"
     Protein         1..268
                     /product="CBEL protein, formerly GP34"
                     /function="elicits defense reactions in host"
     sig_peptide     1..18
     Region          23..55
                     /region_name="CBM_1"
                     /note="Fungal cellulose binding domain; cl02521"
                     /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:O42830"
                     /db_xref="InterPro:IPR000177"
                     /db_xref="InterPro:IPR000254"
                     /db_xref="InterPro:IPR003014"
                     /db_xref="InterPro:IPR003609"
                     /db_xref="UniProtKB/TrEMBL:O42830"
ORIGIN
```