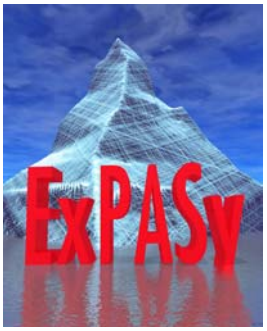


An introduction to biological databases

Marie-Claude.Blatter@isb-sib.ch

What is a database ?

- A collection of
  - structured
  - searchable (index) -> table of contents
  - updated periodically (release) -> new edition
  - cross-referenced (hyperlinks) -> links with other db

  data

- Includes also associated tools (software) necessary for db access/query, db updating, db information insertion, db information deletion....

Why biological databases ?

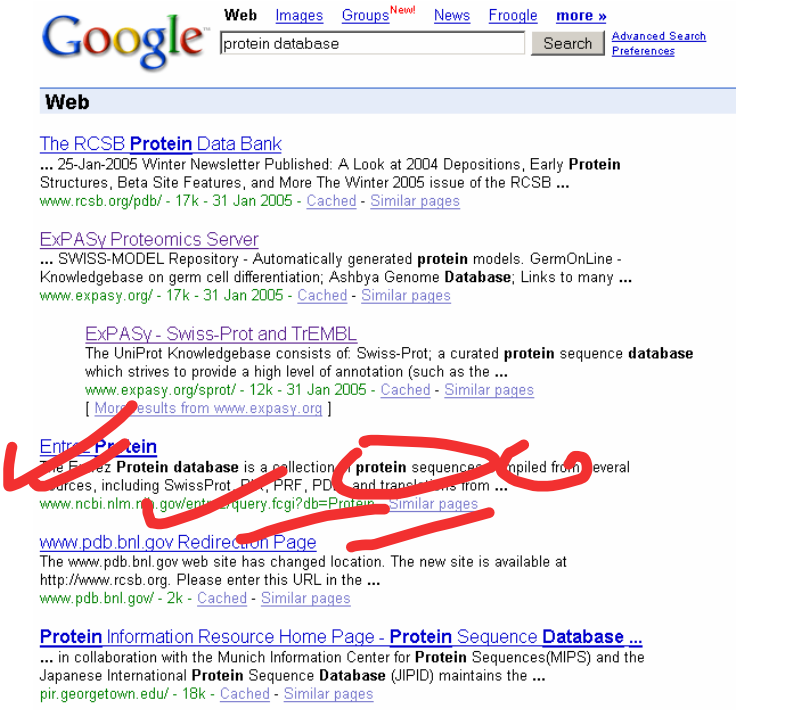
- Exponential growth in biological data.
- Data (genomic sequences, 3D structures, 2D gel analysis, MS analysis, Microarrays ….) are no longer published in a conventional manner, but directly submitted to databases.
- Essential tools for biological research.

Distribution of databases

• Books, articles	1968 -> 1985
• Computer tapes	1982 ->1992
• Floppy disks	1984 -> 1990
• CD-ROM	1989 -> ?
• FTP	1989 -> ?
• On-line services	1982 -> 1994
→ • WWW	1993 -> ?
• DVD	2001 -> ?

Some statistics and remarks

- More than 1000 different 'biological' databases
- Variable size: <100Kb to >10Gb
  - DNA: > 10 Gb
  - Protein: 1 Gb
  - 3D structure: 5 Gb
  - Other: smaller
- Update frequency: daily to annually
- ***How to find them ?***
  - Amos' links: www.expasy.org/alinks.html
  - Biohunt: http://www.expasy.org/BioHunt/
  - Google: http://www.google.com/



EMBnet MCB, feb 2005

Protein Sequences is different than Nucleotide Sequence

The ten important bioinformatics databases *

GenBank/DDJB/EMBL	www.ncbi.nlm.nih.gov	Nucleotide sequences
Ensembl	www.ensembl.org	Human/mouse genome
PubMed	www.ncbi.nlm.nih.gov	Literature references
NR	www.ncbi.nlm.nih.gov	Protein sequences
Swiss-Prot	www.expasy.org	Protein sequences
InterPro	www.ebi.ac.uk	Protein domains
OMIM	www.ncbi.nlm.nih.gov	Genetic diseases
Enzymes	www.expasy.org	Enzymes
PDB	www.rcsb.org/pdb/	Protein structures
KEGG	www.genome.ad.jp	Metabolic pathways

*according to the « Bioinformatics for dummies »

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Ideal minimal content of a sequence database entry

- Sequences !!
- Accession number (AC) (unique identifier)
- Taxonomic data
- References
- ANNOTATION/CURATION
- Keywords
- Cross-references
- Documentation

Sequence Databases: some « technical » definitions

Data storage management:
- flat file: text file, human readable
- relational database (e.g., Oracle, Postgres)
- object oriented database

Sequence format (for BLAST, prediction tools…)
- Fasta, RAW
- GCG
- NBRF/PIR
- MSF….
- standardized format ?

Sequence database : format

SWISS-PROT (protein db) (flat file)

Accession number

Taxonomy

Reference

Annotations (comments)

Cross-references

Keywords

```
ID   EPO_HUMAN      STANDARD;      PRT;   193 AA.
AC   P01588; Q9UHA0; Q9UEZ5; Q9UDZ0;
DT   21-JUL-1986 (Rel. 01, Created)
DT   21-JUL-1986 (Rel. 01, Last sequence update)
DT   20-AUG-2001 (Rel. 40, Last annotation update)
DE   Erythropoietin precursor.
GN   EPO.
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   SEQUENCE FROM N.A.
RX   MEDLINE=85137899; PubMed=3838366;
RA   Jacobs K., Shoemaker C., Rudersdorf R., Neill S.D., Kaufman R.J.,
RA   Mufson A., Seehra J., Jones S.S., Hewick R., Fritsch E.F.,
RA   Kawakita M., Shimizu T., Miyake T.;
RT   "Isolation and characterization of genomic and cDNA clones of human
RT   erythropoietin.";
RL   Nature 313:806-810(1985).
....
CC   -!- FUNCTION: ERYTHROPOIETIN IS THE PRINCIPAL HORMONE INVOLVED IN THE
CC       REGULATION OF ERYTHROCYTE DIFFERENTIATION AND THE MAINTENANCE OF A
CC       PHYSIOLOGICAL LEVEL OF CIRCULATING ERYTHROCYTE MASS.
CC   -!- SUBCELLULAR LOCATION: SECRETED.
CC   -!- TISSUE SPECIFICITY: PRODUCED BY KIDNEY OR LIVER OF ADULT MAMMALS
CC       AND BY LIVER OF FETAL OR NEONATAL MAMMALS.
CC   -!- PHARMACEUTICAL: Available under the names Epogen (Amgen) and
CC       Procrit (Ortho Biotech).
…
DR   EMBL; X02158; CAA26095.1; -.
DR   EMBL; X02157; CAA26094.1; -.
DR   EMBL; M11319; AAA52400.1; -.
DR   EMBL; AF053356; AAC78791.1; -.
DR   EMBL; AF202308; AAF23132.1; -.
DR   EMBL; AF202306; AAF23132.1; JOINED.
....
KW   Erythrocyte maturation; Glycoprotein; Hormone; Signal; Pharmaceutical.
```

Sequence database: format

Annotations (features)

Sequence

```
FT   SIGNAL          1     27
FT   CHAIN          28    193       ERYTHROPOIETIN.
FT   PROPEP        190    193       MAY BE REMOVED IN PROCESSED PROTEIN.
FT   DISULFID       34    188
FT   DISULFID       56     60
FT   CARBOHYD       51     51       N-LINKED (GLCNAC...).
FT   CARBOHYD       65     65       N-LINKED (GLCNAC...).
FT   CARBOHYD      110    110       N-LINKED (GLCNAC...).
FT   CARBOHYD      153    153       O-LINKED (GALNAC...).
FT   VARIANT       131    132       SL -> NF (IN AN HEPATOCELLULAR
FT                                  CARCINOMA).
FT                                  /FTId=VAR_009870.
FT   VARIANT       149    149        P -> Q (IN AN HEPATOCELLULAR CARCINOMA).
FT                                  /FTId=VAR_009871.
FT   CONFLICT       40     40        E -> Q (IN REF. 1; CAA26095).
FT   CONFLICT       85     85        Q -> QQ (IN REF. 5).
FT   CONFLICT      140    140        G -> R (IN REF. 1; CAA26095).
**
**   ##################    INTERNAL SECTION    ##################
**CL 7q22;
SQ   SEQUENCE   193 AA;  21306 MW;  C91F0E4C26A52033 CRC64;
     MGVHECPAWL WLLLSLLSLP LGLPVLGAPP RLICDSRVLE RYLLEAKEAE NITTGCAEHC
     SLNENITVPD TKVNFYAWKR MEVGQQAVEV WQGLALLSEA VLRGQALLVN SSQPWEPLQL
     HVDKAVSGLR SLTTLLRALG AQKEAISPPD AASAAPLRTI TADTFRKLFR VYSNFLRGKL
     KLYTGEACRT GDR
//
```

Sequence database: format

…The fasta format:

```
> My_Sequence_Name
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPD
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```

…The RAW format:

```
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPD
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```


Database 1a: nucleotide sequences

- The 3 main public nucleic acid sequence databases are EMBL (Europe)/GenBank (USA) /DDBJ (Japan)
  « different views of the same data set » within 2 to 3 days (since 1990)
- EMBL: since 1982
- Specialized databases for the different types of RNAs (i.e. tRNA, rRNA, tm RNA, uRNA, etc…)
- 3D structure (DNA and RNA) → PDB
- Others: Aberrant splicing db; Eukaryotic promoter db (EPD); RNA editing sites, Multimedia Telomere Resource ……

Nucleotide and associated topics databases

- EMBL - EMBL Nucleotide sequence db (EBI)
- Genbank - GenBank Nucleotide Sequence db (NCBI)
- DDBJ - DNA Data Bank of Japan
- dbEST - dbEST (Expressed Sequence Tags) db (NCBI)
- dbSTS - dbSTS (Sequence Tagged Sites) db (NCBI)

- NDB - Nucleic Acid Databank (3D structures)
- BNASDB - Nucleic acid structure db from University of Pune

- AsDb - Aberrant Splicing db
- ACUTS - Ancient conserved untranslated DNA sequences db
- Codon Usage Db
- EPD - Eukaryotic Promoter db
- HOVERGEN - Homologous Vertebrate Genes db
- IMGT - ImMunoGeneTics db [Mirror at EBI]
- ISIS - Intron Sequence and Information System
- RDP - Ribosomal db Project
- gRNAs db - Guide RNA db
- PLACE - Plant cis-acting regulatory DNA elements db
- PlantCARE - Plant cis-acting regulatory DNA elements db
- sRNA db - Small RNA db
- ssu rRNA - Small ribosomal subunit db
- lsu rRNA - Large ribosomal subunit db
- 5S rRNA - 5S ribosomal RNA db
- tmRNA Website
- tmRDB - tmRNA dB
- tRNA - tRNA compilation from the University of Bayreuth
- uRNADB - uRNA db
- RNA editing - RNA editing site
- RNAmod db - RNA modification db
- SOS-DGBD - Db of Drosophila DNA sequences annotated with regulatory binding sites
- TelDB - Multimedia Telomere Resource
- TRADAT - TRAnscription Databases and Analysis Tools
- Subviral RNA db - Small circular RNAs db (viroid and viroid-like)

- MPDB - Molecular probe db
- OPD - Oligonucleotide probe db
- VectorDB - Vector sequence db (seems dead!)

Amos'links
http://www.expasy.org/alinks.html#DNA

Real life of a sequence ...

Data not submitted to public databases*, delayed or cancelled...

with or without annotated CDS provided by authors

cDNAs, ESTs, genes, genomes, ...

EMBL, GenBank, DDBJ

CDS
CoDing Sequence
portion of DNA/RNA translated into protein
(from Met to STOP)

Experimentally proved
or derived from gene prediction

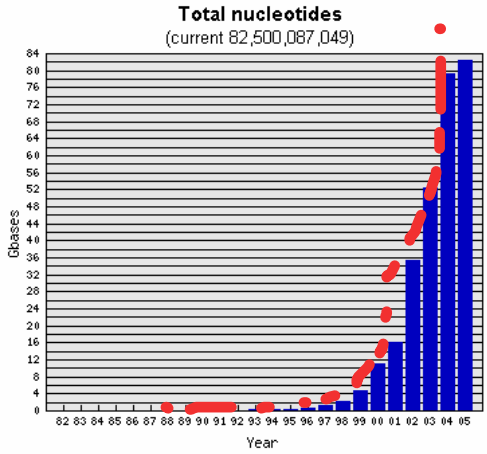
* REMARK: Journals do not accept a paper dealing with a sequence if the EMBL/GenBank/DDBJ AC number is not available

EMBL/GenBank/DDBJ

- Serve as archives
- Contain all public sequences derived from:
  - Genome projects (> 80 % of entries)
  - Sequencing centers (cDNAs, ESTs…)
  - Individual scientists (15 % of entries)
  - Patent offices (i.e. European Patent Office, EPO)

- Currently: 46×10^6 sequences, ~80×10^9 bp;
  Sequences from > 80'000 different species;
  - Contribution: EMBL 10 %; GenBank 73 %; DDBJ 17 %

The tremendous increase in nucleotide sequences



1980: 80 genes fully sequenced !

EMBL STATISTICS

This morning the EMBL Database contained **82,500,087,049** nucleotides in **[illegible]** entries.

Breakdown by entry type:

Entry Type	Entries	Nucleotides
Standard	42,213,672	46,346,877,856
Constructed (CON)	403,272	n/a
Third Party Annotation (TPA)	4,549	332,038,478
Whole Genome Shotgun (WGS)	4,618,027	35,770,720,283

RNA
DNA

Top Organisms

By nucleotide count

Homo sapiens | Mus musculus | Rattus norvegicus | Pan troglodytes
Canis familiaris | Danio rerio | Monodelphis domestica | Bos taurus
Gallus gallus | Zea mays | Other

More than 80'000 species, but...

Human/Mouse/Rat:
Organisms with the highest redundancy !

New projects:
Environmental sequences
(no taxonomic information)

an EMBL entry

DNA (genomic)
or
RNA

```
ID   HSERPG     standard; genomic DNA; HUM; 3398 BP.
XX
AC   X02158;
XX
SV   X02158.1
XX
DT   13-JUN-1985 (Rel. 06, Created)
DT   22-JUN-1993 (Rel. 36, Last updated, Version 2)
XX
DE   Human gene for erythropoietin
XX
KW   erythropoietin; glycoprotein hormone; hormone; signal peptide.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Primates; Catarrhini; Hominidae; Homo.
XX
RN   [1]
RP   1-3398
RX   MEDLINE; 85137899.
RA   Jacobs K., Shoemaker C., Rudersdorf R., Neill S.D., Kaufman R.J.,
RA   Mufson A., Seehra J., Jones S.S., Hewick R., Fritsch E.F., Kawakita M.,
RA   Shimizu T., Miyake T.;
RT   Isolation and characterization of genomic and cDNA clones of human
RT   erythropoietin;
RL   Nature 313:806-810(1985).
XX
DR   GDB; 119110; EPO.
DR   GDB; 119615; TIMP1.
DR   Swiss-Prot; P01588; EPO_HUMAN.
XX
...
```

```
ID   HSERPR     standard; RNA; HUM; 1342 BP.
XX
AC   X02157;
XX
SV   X02157.1
XX
DT   13-JUN-1985 (Rel. 06, Created)
DT   30-MAR-1993 (Rel. 43, Last updated, Version 3)
XX
DE   Human mRNA for fetal erythropoietin
```

keyword

taxonomy

references

Cross-references

```
CC   Data kindly reviewed (24-FEB-1986) by K. Jacobs
FH   Key             Location/Qualifiers
FH
FT   source          1..3398
FT                   /db_xref=taxon:9606
FT                   /organism=Homo sapiens
FT   mRNA            join(397..627,1194..1339,1596..1682,2294..2473,2608..3327)
FT   CDS             join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
FT                   /db_xref=SWISS-PROT:P01588
FT                   /product=erythropoietin
FT                   /protein_id=CAA26095.1
FT                   /translation=MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLLE
FT                   AKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRG
FT                   QALLVNSSQPWEPLQLHVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITAD
FT                   TFRKLFRVYSNFLRGKLKLYTGEACRTGDR
FT   mat_peptide     join(1262..1339,1596..1682,2294..2473,2608..2763)
FT                   /product=erythropoietin
FT   sig_peptide     join(615..627,1194..1261)
FT   exon            397..627
FT                   /number=1
FT   intron          628..1193
FT                   /number=1
FT   exon            1194..1339
FT                   /number=2
FT   intron          1340..1595
FT                   /number=2
FT   exon            1596..1682
FT                   /number=3
FT   intron          1683..2293
FT                   /number=3
FT   exon            2294..2473
FT                   /number=4
FT   intron          2474..2607
FT                   /number=4
FT   exon            2608..3327
FT                   /note=3' untranslated region
FT                   /number=5
XX
SQ   Sequence 3398 BP; 698 A; 1034 C; 991 G; 675 T; 0 other;
     agcttctggg cttccagacc cagctacttt gcggaactca gcaacccagg catctctgag     60
     tctccgccca agaccgggat gccccccagg aggtgtccgg gagcccagcc tttcccagat    120
```

CDS
CoDing Sequence
(proposed by submitters)

Annotation

(Prediction or experimentally determined)

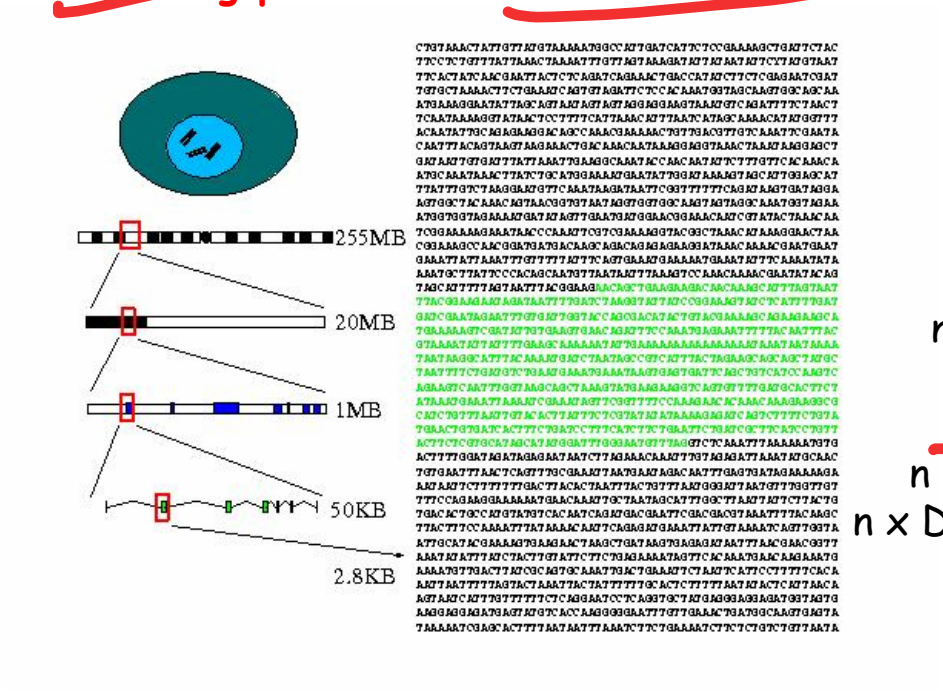
sequence

1: NC_000001. Homo sapiens chro...[gi:51511461]

Click here to see all features and the sequence of this contig record.

```
LOCUS       NC_000001          245522847 bp    DNA     linear   CON 24-AUG-2004
DEFINITION  Homo sapiens chromosome 1, complete sequence.
ACCESSION   NC_000001
VERSION     NC_000001.8  GI:51511461
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 245522847)
  AUTHORS   International Human Genome Sequencing Consortium.
  TITLE     Finishing the euchromatic sequence of the human genome
  JOURNAL   Unpublished (2004)
COMMENT     GENOME ANNOTATION REFSEQ:  Features on this sequence have been
            produced for build 35 version 1 of the NCBI's genome annotation
            [see documentation].
            On Aug 24, 2004 this sequence version replaced gi:42406218.
            The DNA sequence is part of the third release of the finished human
            reference genome. It was assembled from individual clone sequences
            by the Human Genome Sequencing Consortium in consultation with NCBI
            staff.
FEATURES             Location/Qualifiers
     source          1..245522847
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="1"
CONTIG       join(NT_077402.1:1..167280,gap(50000),NT_077911.1:1..40302,
             gap(100000),NT_077912.1:1..153649,gap(50000),NT_077913.3:1..798564,
             gap(50000),NT_077914.2:1..192022,gap(50000),
             NT_004350.17:1..1014565,gap(50000),NT_004321.16:1..1130358,
             gap(60000),NT_004547.16:1..1430159,gap(60000),
             NT_021937.17:1..7618319,gap(50000),NT_077382.3:1..237250,
             gap(50000),NT_004873.16:1..3518496,gap(50000),
             NT_004610.17:1..8601969,gap(50000),NT_037485.3:1..2854727,
             gap(200000),NT_004538.16:1..1064515,gap(5000),
             NT_004511.17:1..16145453,gap(50000),NT_032977.7:1..57683299,
             gap(50000),NT_019273.17:1..16783271,gap(50000),
             NT_086586.1:1..189539,gap(50000),NT_077389.3:1..398739,
             gap(20290000),NT_079485.2:1..164360,gap(300000),
             NT_077932.2:1..127263,gap(50000),NT_077933.1:1..170669,gap(100000),
             NT_004434.17:1..1002899,gap(50000),NT_034398.4:1..281532,
             gap(50000),NT_034400.3:1..1065823,gap(50000),NT_077936.2:1..340928,
             gap(150000),NT_079483.1:1..172789,gap(50000),NT_034401.5:1..220313,
             gap(50000),NT_034403.3:1..309240,gap(50000),NT_086599.1:1..129426,
             gap(50000),NT_004487.17:1..56331646,gap(50000),
             NT_086602.1:1..170924,gap(50000),NT_021877.17:1..16682800,
             gap(50000),NT_077939.1:1..422453,gap(50000),
             NT_004559.12:1..11387711,gap(50000),NT_004836.16:1..13649709,
             gap(50000),NT_032968.7:1..174886,gap(80000))
//
```

The big problem = the redondancy



GSS
HTG
WGS

HTC
n x EST

HUM
n x cDNA
n x DNA (Gene)

...

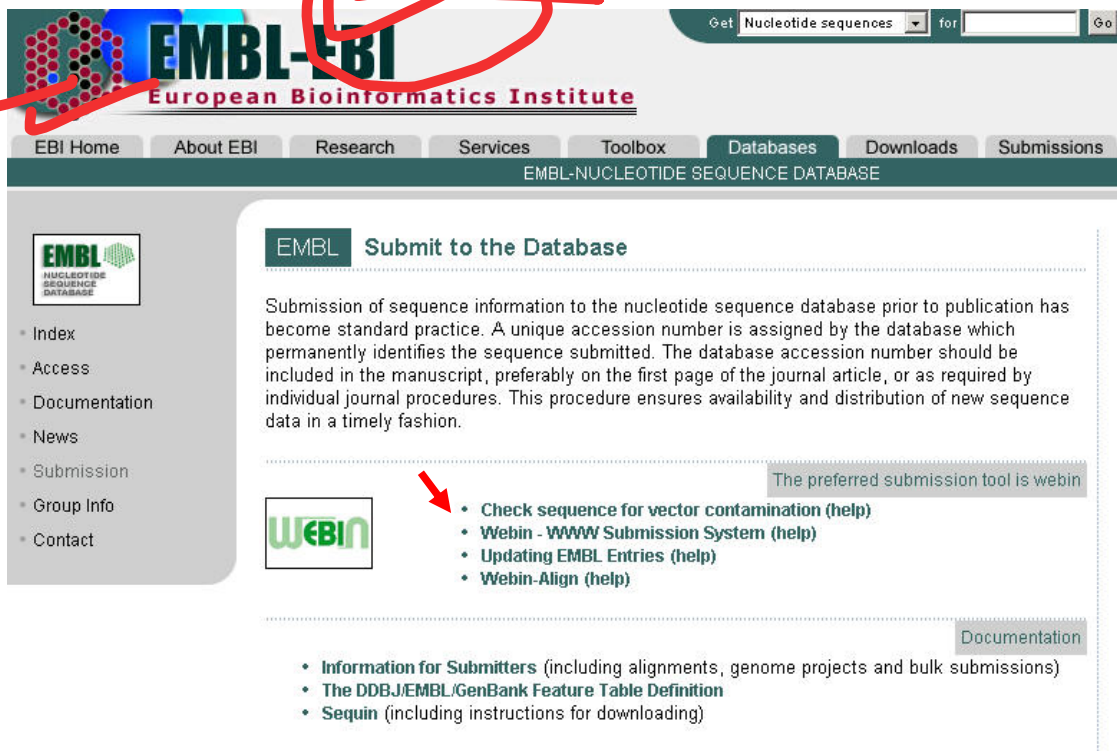
EMBL/GenBank/DDBJ

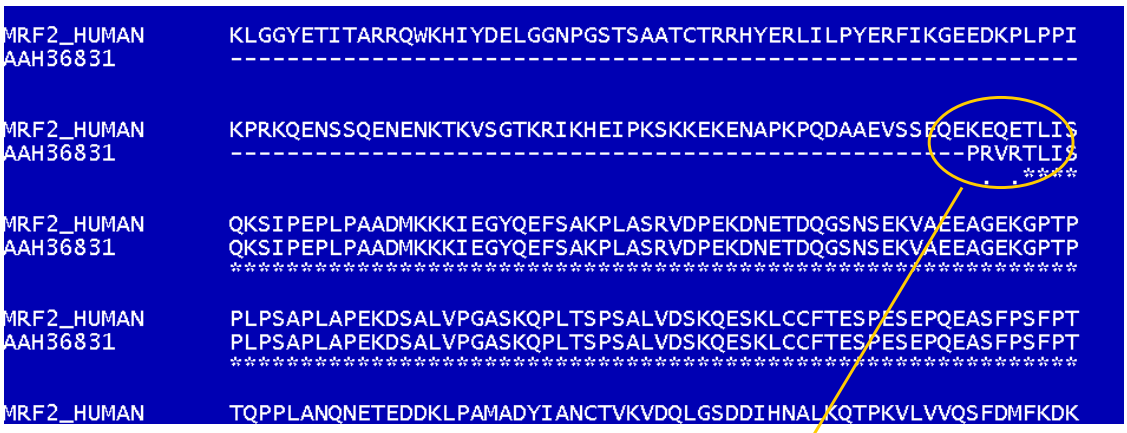
Sort of sequence museum where sequences are preserved for eternity as they were determined, interpreted and published originally by their authors
(primary sequence repository)

The authors have full authority over the content of the entries they submit !
(editorial control of the content belongs to the authors)

(exception: TPA, since january 2003)

Submission: FTP, email, Webin, etc…





Protein sequence derived from the traduction of a vector contamination

EMBL/GenBank/DDBJ

- Unexpected information you can find in these db:

```
FT   source          1..124
FT                   /db_xref="taxon:4097"
FT                   /organelle="plastid:chloroplast"
FT                   /organism="Nicotiana tabacum"
FT                   /isolate="Cuban cahibo cigar, gift from
FT                    President Fidel Castro"
```

- Or:

```
FT   source          1..17084
FT                   /chromosome="complete mitochondrial genome"
FT                   /db_xref="taxon:9267"
FT                   /organelle="mitochondrion"
FT                   /organism="Didelphis virginiana"
FT                   /dev_stage="adult"
FT                   /isolate="fresh road killed individual"
FT                   /tissue_type="liver"
```

```
FT  CDS          complement(45959..47332)
FT               /db_xref="SPTREMBL:Q9UZ71"
FT               /note="PAB2386"
FT               /transl_table=11
FT               /product="4-AMINOBUTYRATE qui se dilate AMINOTRANSFERASE
FT               (EC 2.6.1.19)"
FT               /protein_id="CAB50188.1"
FT               /translation="MDYPRIVVNPPGPKAKELIEREKRVLSTGIGVKLFPLVPKRGFGP
FT               FIEDVDGNVFIDFLAGAAAASTGYSHPKLVKAVKEQVELIQHSMIGYTHSERAIRVAEK
FT               LVKISPIKNSKVLFGLSGSDAVDMAIKVSKFSTRRPWILAFIGAYHGQTLGATSVASFQ
FT               VSQKRGYSPLMPNVFWVPYPNPYRNPWGINGYEEPQELVNRVVEYLEDYVFSHVVPPDE
FT               VAAFFAEPIQGDAGIVVPPENFFKELKKLLDEHGILLVMDEVQTGIGRTGKWFASEWFE
FT               VKPDMIIFGKGVASGMGLSGVIGREDIMDITSGSALLTPAANPVISAAADATLEIIEEE
FT
FT
```

NiceProt View of TrEMBL: Q9UZ71

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features]

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

Entry information	
Entry name	Q9UZ71
Primary accession number	Q9UZ71
Secondary accession numbers	None
Entered in TrEMBL in	Release 13, May 2000
Sequence was last modified in	Release 13, May 2000
Annotations were last modified in	Release 26, March 2004
Name and origin of the protein	
Protein name	4-aminobutyrate QUI SE DILATE aminotransferase
Synonym	EC 2.6.1.19
Gene name	PYRAB12830 or PAB2386
From	Pyrococcus abyssi [TaxID: 29292]
Taxonomy	Archaea; Euryarchaeota; Thermococci; Thermococcales; Thermococcaceae; Pyrococcus.

The second generation of nucleotide sequence databases

Gene-centric databases

All the sequence information relevant to a given gene is made accessible at once

i.e. Locus Link/RefSeq

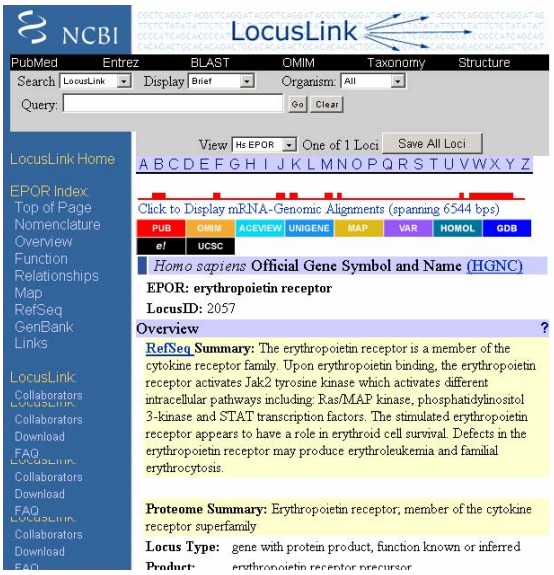
Genome-centric databases

Information about gene sequence, relative position, strand orientation, biochemical functions…

Information management systems that are able to connect specialized sequence collection and browsing tools

i.e. Ensembl, TIGR

Gene-centric databases



New: Replaced by « Entrez Gene » on March 1, 2005

~~LocusLink~~ is tighly linked to RefSeq
(« interdependent curated resources »)
Nucl. Ac. Res., 29, 137-140(2001)

Links to the RefSeq database:
« Reference Sequences»
- for RNA (NM_)
- for genomic (NT_)
- for protein (NP_)

Links to all the sequences found in EMBL/GenBank/DDBJ corresponding to this gene

The corresponding RefSeq entry for the mRNA

NCBI Reference Sequence
http://www.ncbi.nlm.nih.gov/RefSeq/

☐ **1:** NM_000121. Homo sapiens eryt...[gi:4557561]

```
LOCUS       NM_000121               1849 bp    mRNA    linear   PRI 23-AUG-2004
DEFINITION  Homo sapiens erythropoietin receptor (EPOR), mRNA.
ACCESSION   NM_000121
VERSION     NM_000121.2  GI:4557561
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1849)
  AUTHORS   Acs,G., Xu,X., Chu,C., Acs,P. and Verma,A.
  TITLE     Prognostic significance of erythropoietin expression in human
            endometrial carcinoma
  JOURNAL   Cancer 100 (11), 2376-2386 (2004)
   PUBMED   15160341
  REMARK    GeneRIF: Increased erythropoietin receptor expression is associated
            with advanced-stage disease, lymphovascular invasion, lymph node
            metastasis of endometrial carcinoma
REFERENCE   2  (bases 1 to 1849)
  AUTHORS   Eccles,T.G., Patel,A., Verma,A., Nicholson,D., Lukes,Y.,
            Tuttle,R.M. and Francis,G.L.
  TITLE     Erythropoietin and the erythropoietin receptor are expressed by
            papillary thyroid carcinoma from children and adolescents.
            Expression of erythropoietin receptor might be a favorable
            prognostic indicator
  JOURNAL   Ann. Clin. Lab. Sci. 33 (4), 411-422 (2003)
   PUBMED   14584755
  REMARK    GeneRIF: Majority of papillary thyroid carcinomas (PTC) from
            children and adolescents express EPO-R, a finding associated with
            favorable prognostic indicators and a lower risk of recurrence.
REFERENCE   3  (bases 1 to 1849)
  AUTHORS   Arcasoy,M.O., Jiang,X. and Haroon,Z.A.
  TITLE     Expression of erythropoietin receptor splice variants in human
            cancer
  JOURNAL   Biochem. Biophys. Res. Commun. 307 (4), 999-1007 (2003)
   PUBMED   12878211
  REMARK    GeneRIF: Expression of erythropoietin receptor splice variants in
            tumor cell lines.
REFERENCE   4  (bases 1 to 1849)
  AUTHORS   Matsuzaki,S., Canis,M., Yokomizo,R., Yaegashi,N., Bruhat,M.A. and
            Okamura,K.
  TITLE     Expression of erythropoietin and erythropoietin receptor in
            peritoneal endometriosis
  JOURNAL   Hum. Reprod. 18 (1), 152-156 (2003)
  12525458
  GeneRIF: Epo and Epo-R localized within glandular epithelial cells
  in both peritoneal endometriosis and eutopic endometrium. Epo-R
  expression lower in black peritoneal lesions.
5  (bases 1 to 1849)
```

RefSeq

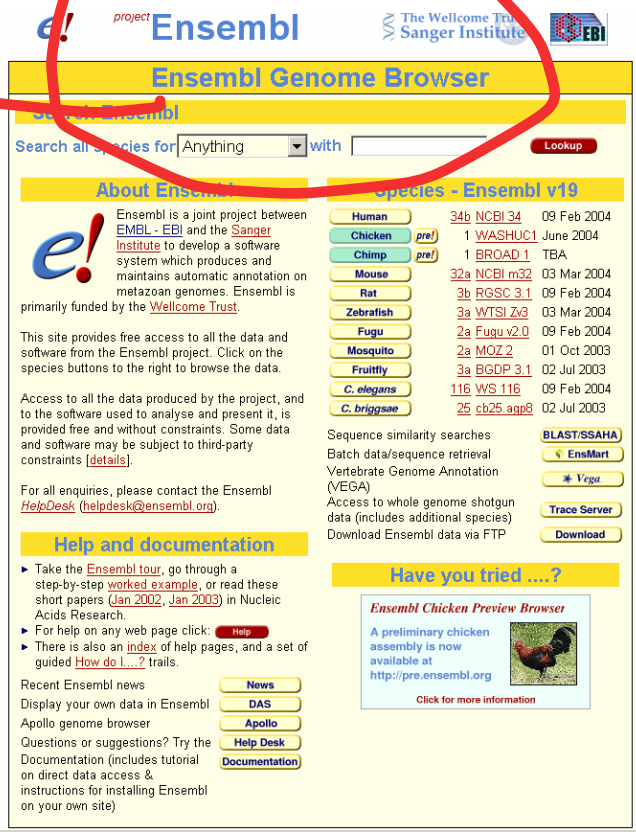
```
  JOURNAL   Blood 76 (1), 24-30 (1990)
    PUBMED  2162695
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from M60459.1 and M76595.1.
            On Apr 1, 1999 this sequence version replaced gi:4503590.

            Summary: The erythropoietin receptor is a member of the cytokine
            receptor family.
            receptor activat
            intracellular pa
            phosphatidylinos
            stimulated eryth
            erythroid cell s
            produce erythrol
            COMPLETENESS: co
FEATURES             Locatio
     source          1..1849
                     /organi
                     /mol_ty
                     /db_xre
                     /chromo
                     /map="1
     gene            1..1849
                     /gene="
                     /db_xre
                     /db_xre
                     /db_xre
     CDS             137..16
                     /gene="EPOR"
                     /note="go_component: integral to plasma membrane [goid
                     0005887] [evidence TAS] [pmid 1848667];
                     go_function: erythropoietin receptor activity [goid
                     0004900] [evidence TAS] [pmid 8506290];
                     go_function: guanyl-nucleotide exchange factor activity
                     [goid 0005085] [evidence IEA];
                     go_process: signal transduction [goid 0007165] [evidence
                     NR] [pmid 1660606];
                     go_process: small GTPase mediated signal transduction
                     [goid 0007264] [evidence IEA]"
                     /codon_start=1
                     /product="erythropoietin receptor precursor"
                     /protein_id="NP_000112.1"
                     /db_xref="GI:4503591"
                     /db_xref="GeneID:2057"
                     /db_xref="LocusID:2057"
                     /db_xref="MIM:133171"
                     /translation="MDHLGASLWPQVGSLCLLLAGAAWAPPPNLPDPKFESKAALLAA
                     RGPEELLCFTERLEDLVCFWEEAASAGVGPGNYSFSYQLEDEPWKLCRLHQAPTARGA
                     VRFWCSLPTADTSSFVPLELRVTAASGAPRYHRVIHINEVVLLDAPVGLVARLADESG
                     HVVLRWLPPPETPMTSHIRYEVDVSAGNGAGSVQRVEILEGRTECVLSNLRGRTRYTF
                     AVRARMAEPSFGGFWSAWSEPVSLLTPSDLDPLILTLSLILVVILVLLTVLALLSHRR
                     ALKQKIWPGIPSPEKFEGLFTTHKGNFQLWLYQNDGCLWWSPCTPFTEDPPASLEVL
                     SERCWGTMQAVEPGTDDEGPLLEPVGSEHAQDTYLVLDKWLLPRNPPSEDLPGPGGSV
                     DIVAMDEGSEAASSCSSALASKPSPEGASAASFEYTILDPSSQLLRPWTLCPELPPTPP
                     HLKYLYLVVSDSGISTDYSSGDSQGAQGGLSDGPYSNPYENSLIPAAEPLPPSYVACS
                     "
     sig_peptide     137..208
                     /gene="EPOR"
     mat_peptide     209..1660
                     /gene="EPOR"
                     /product="erythropoietin receptor"
ORIGIN
        1 acttagaggc gcctggtcgg gaagggcctg gtcagctgcg tccggcgggag gcagctgctg
       61 acccagctgt ggactgtgcc gggggcgggg gacggagggg caggagccct gggctccccg
      121 tggcgggggc tgtatcatgg accacctcgg ggcgtcccte tggccccagg tcggctccct
      181 ttgtctcctg ctcgctgggg ccgcctgggc gccccccgct aacctcccgg acccaagtt
      241 cgaggcgcaaa gcggcctgc tggcggccag gggccccgaa gagcttctgt gttcagcga
      301 gcggttggag gacttggtgt gtttctggga ggaagcggcg agcgctgggg tgggccccgg
      361 caactacagc ttctcctacc agctcgagga tgagccatgg aagctgtgtc gcctgcacca
```

Working with whole genome databases:

Genome-centric databases

« Browsing resources »

Remark: Genome-centric databases give usually access to several genomes, but some are « specialized » in particular organisms, i.e. TIGR: bacteria and plants



Ensembl provides a bioinformatics framework to organise biology around the sequences of large genomes.

Available now are: human, mouse, rat, fugu, zebrafish, mosquito, Drosophila, C. elegans, and C. briggsae, chicken…

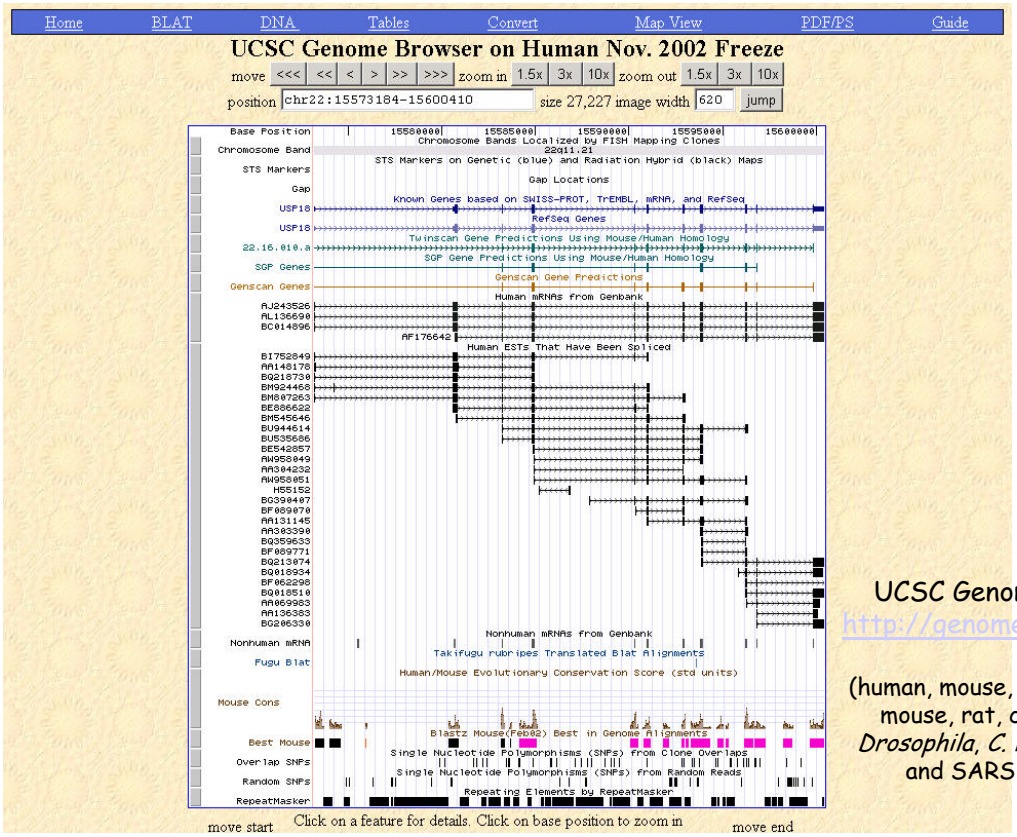
http://www.ensembl.org/

Ensembl/martview: example of queries

- Retrieve all mouse homologues of human disease genes containing transmembrane domains located between 1p22 and 1q22

- Retrieve the sequences 5kb upstream of all human « known » genes from chromosome 6

....



UCSC Genome Browser:
http://genome.cse.ucsc.edu/

(human, mouse, rat chimpanzee, mouse, rat, chicken, *Fugu*, *Drosophila*, *C. briggsae*, yeast, and SARS genomes.)

..and plants

http://www.tigr.org/tdb/

Database 1b: protein sequences

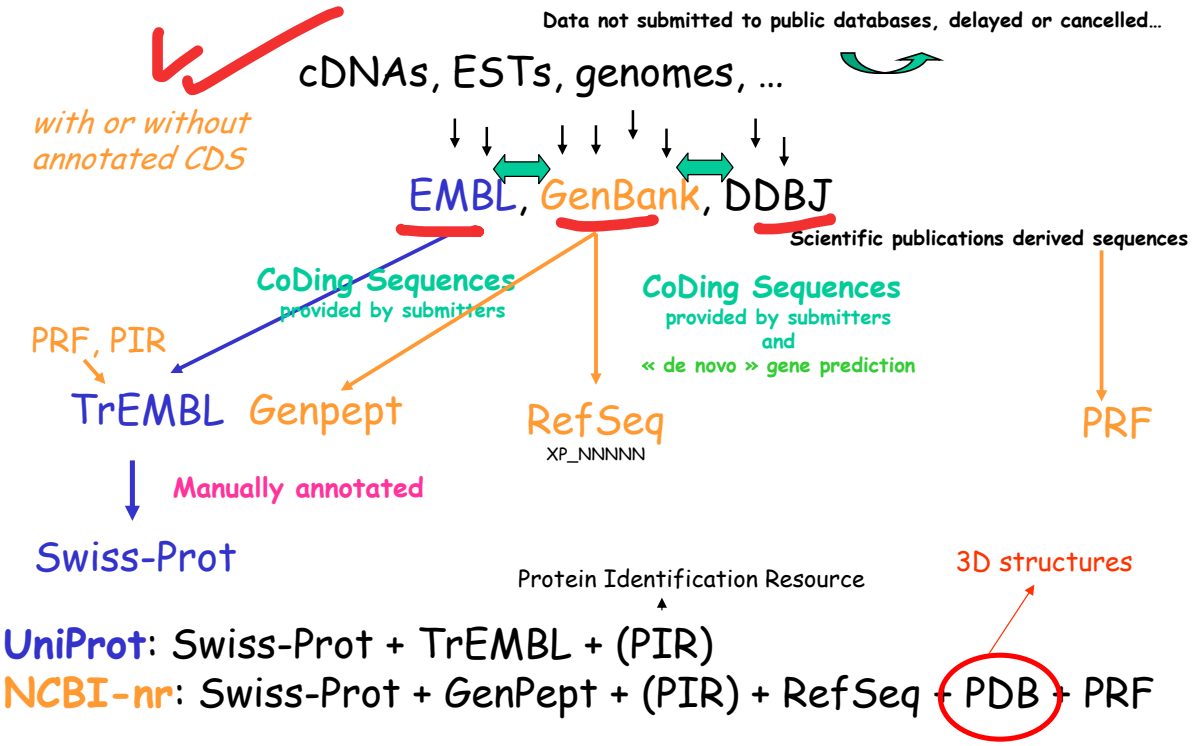
- **SWISS-PROT**: created in 1986 (A.Bairoch) http://www.expasy.org/sprot/
- **TrEMBL**: created in 1996; complement to SWISS-PROT; derived from EMBL CDS translations (« proteomic » version of EMBL)
- **(PIR-PSD**: Protein Information Resources) http://pir.georgetown.edu/

- **Genpept**: « proteomic » version of GenBank (~TrEMBL)
- **RefSeq** (NP_)
- **PRF**

- Many specialized protein databases for specific families or groups of proteins.

  Examples: **AMSDb** (antibacterial peptides), **GPCRDB** (7 TM receptors), **IMGT** (immune system) **YPD** (Yeast) etc.

Real life of a protein sequence …



Protein sequence databases

The UniProt pathway

a central ressource for protein sequences and function…

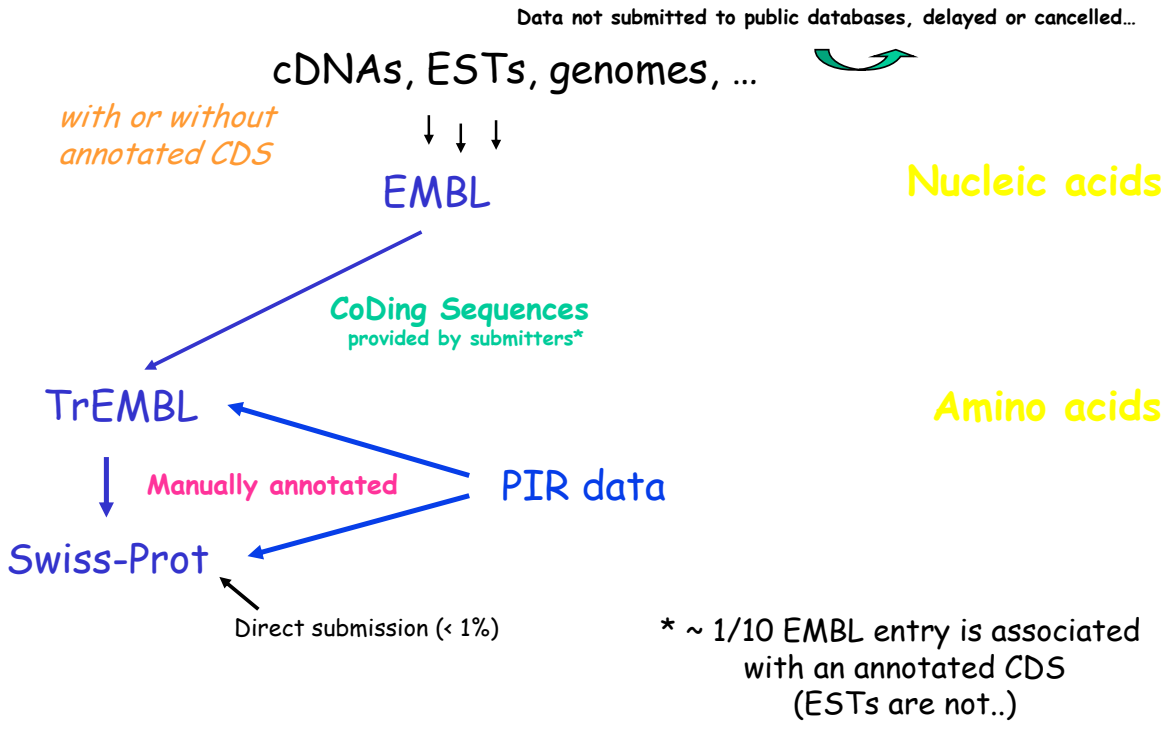
Since december 15, 2003

Swiss-Prot and TrEMBL constitute the

Knowledgebase

(integration of the PIR data)
(Protein Information Ressource)

Real life of a protein sequence …



SWISS-PROT
Protein knowledgebase
TrEMBL
Computer-annotated supplement to SWISS-PROT

UniProt

-> give access to

all known* protein sequences

* submitted to the public databases (EMBL, GenBank, DDJB, Swiss-Prot)

EMBL

CDS

TrEMBL

Swiss-Prot

EMBL

CDS

TrEMBL

Once in Swiss-Prot,
no more in TrEMBL
-> Minimal redundancy

Annotation of sequence differences
(conflicts, variants, splicing…)

Average of 4.2 independent sequence reports
for each human protein

Swiss-Prot

EMBL

CDS

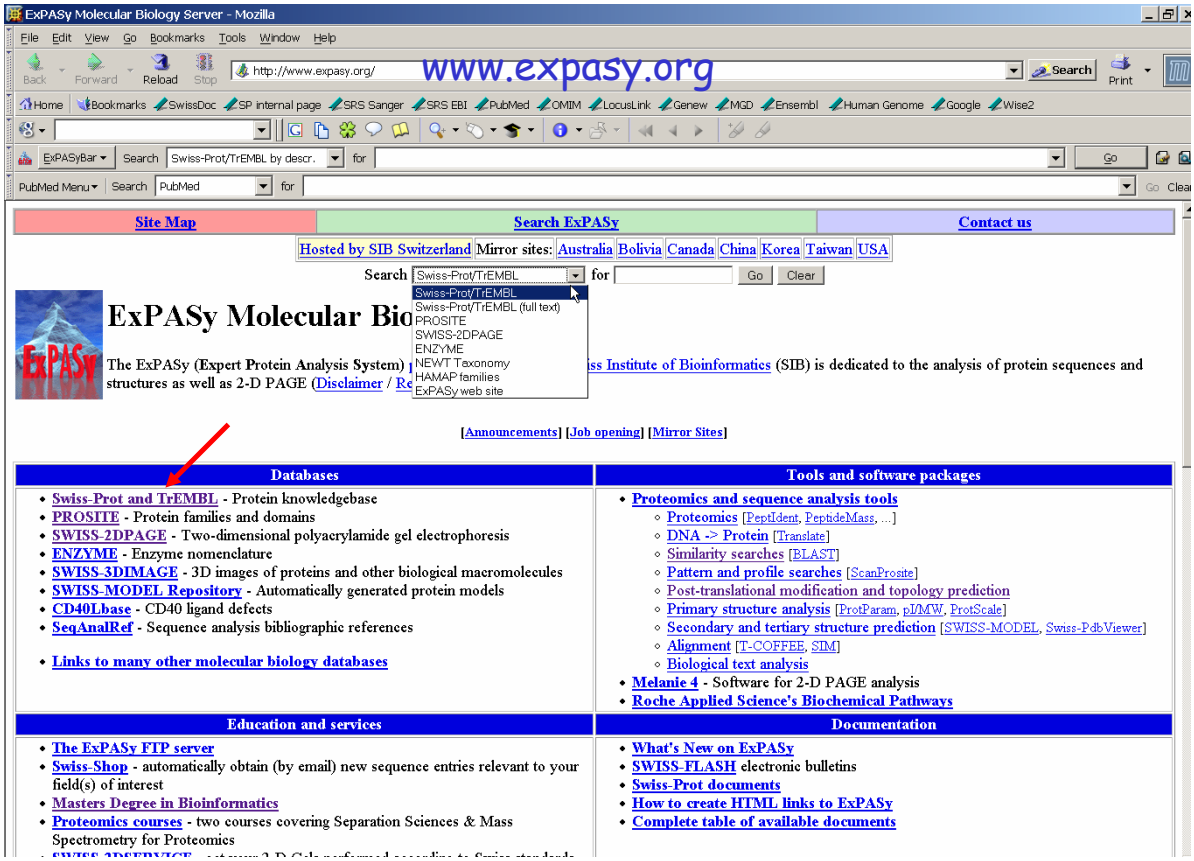
TrEMBL

Swiss-Prot

Up-to-date sources:

Swiss-Prot -> **ExPASy**
Since 1986 (www.expasy.org);

TrEMBL -> **EBI** (European Bioinformatics Institute)
Since 1996 (www.ebi.ac.uk/trembl/).



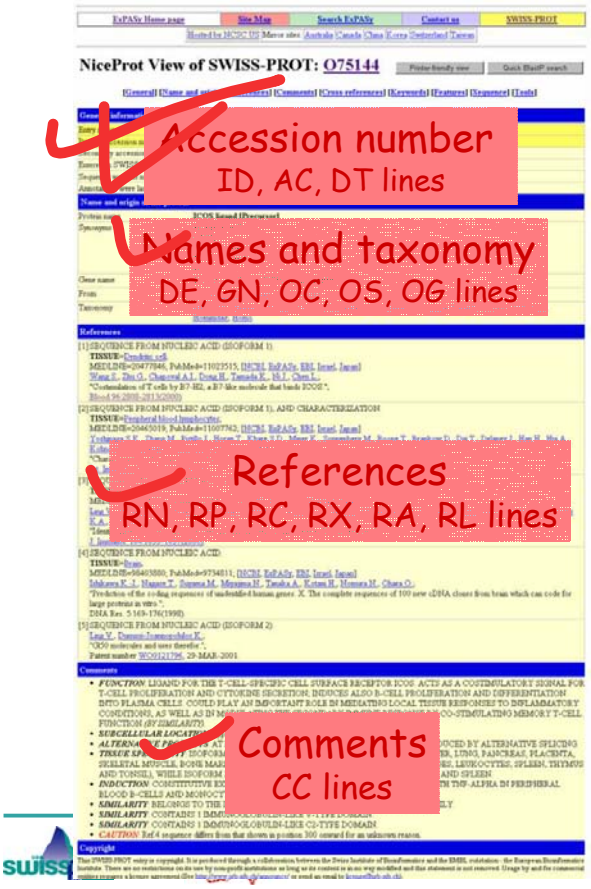
ExPASy

EBI

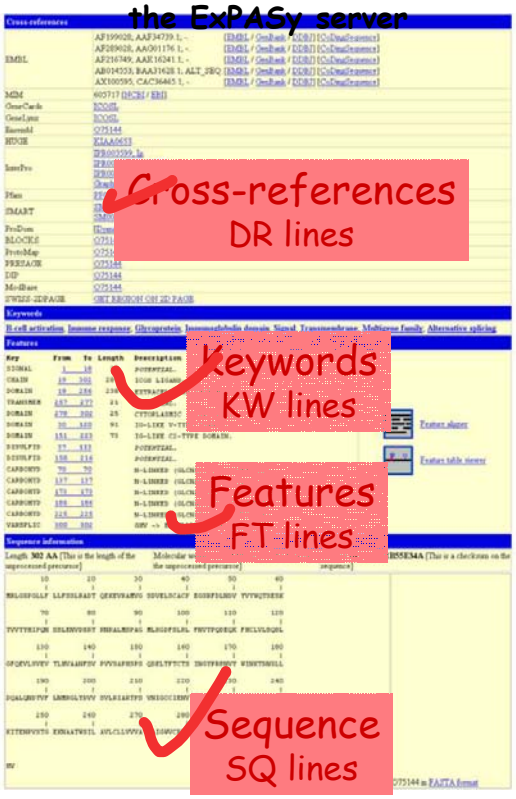
NCBI

In a Swiss-Prot entry, you can expect to find:

- All the **names** of a given protein (and of its gene);
- Its biological origin with links to the taxonomic databases;
- A selection of **references**;
- A summary of what is known about the protein: function, alternative products, PTM, tissue expression, disease, etc.…;
- Numerous **cross-references;**
- Selected keywords;
- A description of important sequence features: domains, PTMs, variations, etc.;
- A (often corrected) protein **sequence** and the description of various isoforms/variants.



View « by default » on the ExPASy server



MCB, feb 2005

Sequence quality

-> 1 gene / 1 specie = 1 Swiss-Prot entry

For human: ~ 4,2 different independent sequence reports /gene

```
SEQ1_HUMAN      MNPESSIFIEDYLKYFQDQVSREMLLQLLTDDEAWNGFILAAELPRDEADELRKALNKLA
SEQ2_HUMAN      ------------------------MLLRLLTDDEAWNGFV-AAELPRDEADELRKALNKLA
                                        ***:************: **********************

SEQ1_HUMAN      SHMVMKDKNR-----------SKEFPRLKRELEDHIRKLRALAEEVEUVHRGTTIANVVS
SEQ2_HUMAN      SHMVMKDKNRSDKDTQHRQWHSKEFPRLKRELEDHIRKLRALAEEVE------------
                **********           ********************************
```

Same gene ?

Polymorphisms ?

Alternative splicing ?

Alternative initiation ?

RNA editing ?

Usage of an alternative promoter ?

Selenocystein ?

Fragment ?

Sequencing errors ?

-> Identification and annotation of all sequence differences

Annotation (Comment lines)

- Function(s) and role(s); enzymes:
  a. Catalytic activity (if EC number)
  b. Cofactor
  c. Enzyme regulation
  d. Pathway
- Subunit (Protein/protein interactions)
- Subcellular location
- Alternative products (alt. splicing, alt. initiation, RNA editing)
- Tissue specificity (Nothern and Western results)
- Developmental stage
- Induction
- Domain
- Post-translational modifications (PTM)
- Mass spectrometry
- Polymorphisms
- Disease
- Pharmaceutical
- Miscellaneous
- Similarities
- Caution
- Database (specialized cross-references)

Annotation/Curation (Comment lines)

Information is derived from:

- **Publications;** currently Swiss-Prot cites 1'500 different journals. 106 journals are cited more than 100 times.
- **Databases;**
- **Personal communication;**
- **Prediction;**
- **Brain storming…**

Experimental qualifiers:

« - »: experimentally proved;
« By similarity »: experimentally proved in an ortholog or in another member of the family;
« Probable »: not proved, but realistic;
« Potential »: predicted (bioinformatic tools).

Comments — ICOL_HUMAN, O75144

- ***FUNCTION***: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-STIMULATING MEMORY T-CELL FUNCTION *(BY SIMILARITY)*.
- ***SUBCELLULAR LOCATION***: Type I membrane protein *(By similarity)*.
- ***ALTERNATIVE PRODUCTS***:
  - **Alternative splicing [2 named forms]**
    Comment: Additional isoforms seem to exist.

Name	1
Isoform ID	O75144-1
This is the isoform sequence displayed in this entry.	

Name	2
Isoform ID	O75144-2
Features which should be applied to build the isoform sequence: VSP_002520.	

- ***TISSUE SPECIFICITY***: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART, KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES, SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN LYMPH NODES, LEUKOCYTES AND SPLEEN.
- ***INDUCTION***: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
- ***SIMILARITY***: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG SUBFAMILY.
- ***SIMILARITY***: Contains 1 immunoglobulin-like V-type domain.
- ***SIMILARITY***: Contains 1 immunoglobulin-like C2-type domain.
- ***CAUTION***: Ref.4 sequence differs from that shown in position 300 onward for an unknown reason.

Experimental qualifiers:

« - »: experimentally proved;
« By similarity »: experimentally proved in an ortholog or in another member of the family;
« Probable »: not proved but realistic;
« Potential »: predicted (bioinformatic tools).

Comments — BRH2_HUMAN, Q9NY43

- ***FUNCTION***: Potential regulator of neural basic helix-loop-helix genes *(By similarity)*.
- ***SUBCELLULAR LOCATION***: Nuclear *(Probable)*.
- ***SIMILARITY***: BELONGS TO THE BAR HOMEOBOX FAMILY.

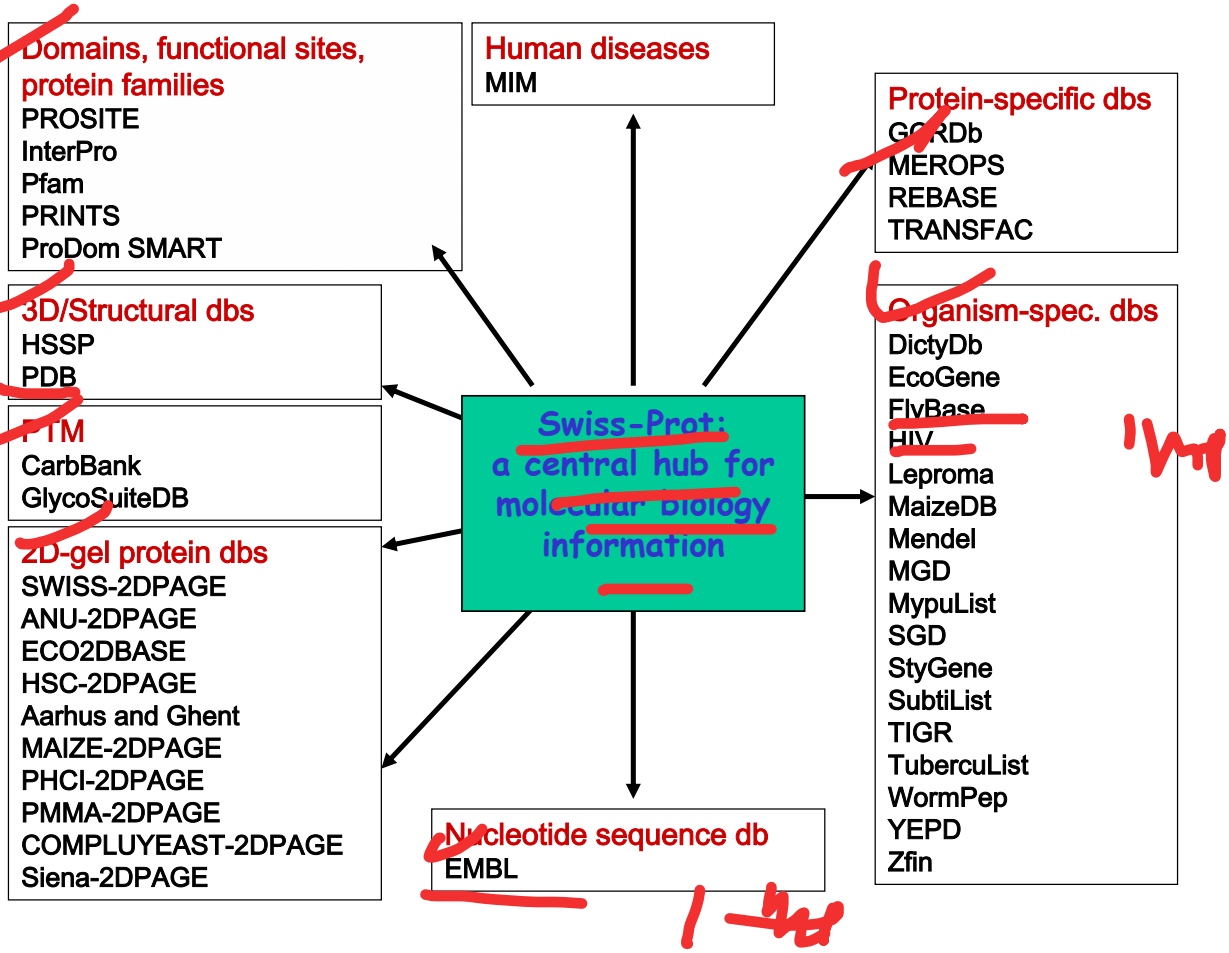
Cross-references

- Explicit links to about 50 databases;
- Implicit X-references to 30 additional db *added by the ExPASy servers on the WWW* (such as GenBank, Ensembl, ...)

  => links to more than 80 databases from the ExPASy servers
- Currently 1.5×10^6 cross-references in Swiss-Prot

-> *Connected with practically all the databases indexed under SRS.*

Gasteiger et al., Curr. Issues Mol. Biol. (2001), 3(3): 47-55



Cross-references

ADN
(Index of low redundancy)

3D

genomic

Examples of implicit links to GenBank and DDBJ added 'on the fly' by the ExPASy server

Cross-references	ICE8_HUMAN Q14790
EMBL	X98172; CAA66853.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] X98173; CAA66854.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] X98174; CAA66855.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] X98175; CAA66856.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] X98176; CAA66857.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] X98177; CAA66858.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoDingSequence] X98178; CAA66859.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoDingSequence] U58143; AAC50602.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] U60520; AAC50645.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF102146; AAD24962.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF102139; AAD24962.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF102140; AAD24962.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF102141; AAD24962.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF102142; AAD24962.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF102143; AAD24962.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF102144; AAD24962.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF102145; AAD24962.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF009620; AAB70913.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AB038985; BAB32555.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AB038982; BAB32555.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AB038983; BAB32555.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AB038984; BAB32555.1; JOINED [EMBL / GenBank / DDBJ] [CoDingSequence] AF380342; AAK57437.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF422925; AAL87628.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF422926; AAL87629.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF422927; AAL87630.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF422928; AAL87631.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] AF422929; AAL87632.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] BC028223; AAH28223.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PDB	1QDU; 10-JUL-00 [ExPASy / RCSB]
MEROPS	C14.009; -.
Genew	HGNC:1509; CASP8.
CleanEx	HGNC:1509; CASP8.
MIM	601763 [NCBI / EBI] 607271 [NCBI / EBI]
GeneCards	CASP8.
GeneLynx	CASP8; Homo sapiens.
SOURCE	CASP8; Homo sapiens.
Ensembl	Q14790; Homo sapiens. [Entry / Contig view]

FT lines - Feature table = Sequence description

Data derived from:
- Publications;
- Databases;
- Personal communication;
- Prediction.

General topology

Features	ICOL_HUMAN, O75144			
Key	From	To	Length	Description
SIGNAL	1	18	18	*POTENTIAL*.
CHAIN	19	302	284	ICOS LIGAND.
DOMAIN	19	256	238	EXTRACELLULAR (*POTENTIAL*).
TRANSMEM	257	277	21	*POTENTIAL*.
DOMAIN	278	302	25	CYTOPLASMIC (*POTENTIAL*).
DOMAIN	30	120	91	IG-LIKE V-TYPE DOMAIN.
DOMAIN	151	223	73	IG-LIKE C2-TYPE DOMAIN.
DISULFID	37	113		*POTENTIAL*.
DISULFID	158	216		*POTENTIAL*.
CARBOHYD	70	70		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	137	137		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	173	173		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	186	186		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	225	225		N-LINKED (GLCNAC...) (*POTENTIAL*).
VARSPLIC	300	302		GHV -> ESWNLLLLLS (IN ISOFORM 2).

Feature aligner

Feature table viewer

Sequence description:

- Derived from:
  - Publications;
  - Databases;
  - Personal communication;
  - Prediction.

Features — ICOL_HUMAN, O75144

Key	From	To	Length	Description
SIGNAL	1	18	18	*POTENTIAL.*
CHAIN	19	302	284	ICOS LIGAND.
DOMAIN	19	256	238	EXTRACELLULAR (*POTENTIAL*).
TRANSMEM	257	277	21	*POTENTIAL.*
DOMAIN	278	302	25	CYTOPLASMIC (*POTENTIAL*).
DOMAIN	30	120	91	IG-LIKE V-TYPE DOMAIN.
DOMAIN	151	223	73	IG-LIKE C2-TYPE DOMAIN.
DISULFID	37	113		*POTENTIAL.*
DISULFID	158	216		*POTENTIAL.*
CARBOHYD	70	70		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	137	137		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	173	173		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	186	186		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	225	225		N-LINKED (GLCNAC...) (*POTENTIAL*).
VARSPLIC	300	302		GHV -> ESWNLLLLLS (IN ISOFORM 2).

PTM


Feature aligner


Feature table viewer

BRC2_HUMAN, P51587

VARIANT	2466	2466	*	A -> V (POLYMORPHISM; WAS ORIGINALLY THOUGHT TO BE LINKED TO OVARIAN CANCER).	VAR_008785
VARIANT	2480	2480	*	L -> V.	VAR_008786
VARIANT	2490	2490	*	I -> T.	VAR_008787
VARIANT	2502	2502	*	R -> H (IN OVARIAN CANCER; COULD BE A POLYMORPHISM).	VAR_008788
VARIANT	2515	2515	*	T -> I (IN BREAST CANCER; COULD BE A POLYMORPHISM).	VAR_008789
VARIANT	2787	2787	*	R -> H (IN OVARIAN CANCER; SOMATIC MUTATION).	VAR_008790
VARIANT	2944	2944	*	I -> F.	VAR_008791
VARIANT	2951	2951	*	A -> T.	VAR_008792
VARIANT	2969	2969	*	V -> M.	VAR_008793
VARIANT	3095	3095	*	D -> E.	VAR_005108
VARIANT	3098	3098	*	Y -> H (IN OVARIAN CANCER; COULD BE A POLYMORPHISM).	VAR_008794
VARIANT	3103	3103	*	I -> M (IN MELANOMA).	VAR_005109
VARIANT	3118	3118	*	M -> T (IN BREAST CANCER).	VAR_005110
VARIANT	3257	3257	*	K -> R.	VAR_008795
VARIANT	3276	3276	*	R -> S.	VAR_008796
VARIANT	3357	3357	*	T -> R (IN BREAST CANCER).	VAR_005111
VARIANT	3412	3412	*	I -> V (POLYMORPHISM; WAS ORIGINALLY THOUGHT TO BE ASSOCIATED WITH BREAST CANCER; dbSNP:1801426) [NCBI/Ensembl].	VAR_005112
CONFLICT	599	599		S -> F (IN REF. 2).	
CONFLICT	758	758		S -> N (IN REF. 1).	
CONFLICT	1761	1762		GY -> RI (IN REF. 1).	
CONFLICT	1767	1767		K -> N (IN REF. 1).	
CONFLICT	2536	2536		S -> P (IN REF. 3).	
CONFLICT	3216	3216		L -> LVS (IN REF. 3).	

Polymorphisms

Differences between the sequence shown and other submitted sequences

Sequence description:

Derived from:

- Publications;
- Databases;
- Personal communication;
- Prediction.

Features — **ICOL_HUMAN, O75144**

Key	From	To	Length	Description
SIGNAL	1	18	18	*POTENTIAL*.
CHAIN	19	302	284	ICOS LIGAND.
DOMAIN	19	256	238	EXTRACELLULAR (*POTENTIAL*).
TRANSMEM	257	277	21	*POTENTIAL*.
DOMAIN	278	302	25	CYTOPLASMIC (*POTENTIAL*).
DOMAIN	30	120	91	IG-LIKE V-TYPE DOMAIN.
DOMAIN	151	223	73	IG-LIKE C2-TYPE DOMAIN.
DISULFID	37	113		*POTENTIAL*.
DISULFID	158	216		*POTENTIAL*.
CARBOHYD	70	70		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	137	137		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	173	173		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	186	186		N-LINKED (GLCNAC...) (*POTENTIAL*).
CARBOHYD	225	225		N-LINKED (GLCNAC...) (*POTENTIAL*).
VARSPLIC	300	302		GHV -> ESWNLLLLLS (IN ISOFORM 2).

Feature aligner

Feature table viewer

Alternative splicing

Alternative Splicing of Swiss-Prot entry: O75144

Name	2
Isoform ID	O75144-2
Features which should be applied to build the isoform sequence:	V

All the alternatively spliced sequences are available for BLAST searches and proteomic tools at the ExPASy server

Sequence information

Length: 309 AA

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MRLGSPGLLF LLFSSLRADT QEKEVRAMVG SDVELSCACP EGSRFDLNDV YVYWQTSESK
        70         80         90        100        110        120
         |          |          |          |          |          |
TVVTYHIPQN SSLENVDSRY RNRALMSPAG MLRGDFSLRL FNVTPQDEQK FHCLVLSQSL
       130        140        150        160        170        180
         |          |          |          |          |          |
GFQEVLSVEV TLHVAANFSV PVVSAPHSPS QDELTFTCTS INGYPRPNVY WINKTDNSLL
       190        200        210        220        230        240
         |          |          |          |          |          |
DQALQNDTVF LNMRGLYDVV SVLRIARTPS VNIGCCIENV LLQQNLTVGS QTGNDIGERD
       250        260        270        280        290        300
         |          |          |          |          |          |
KITENPVSTG EKNAATWSIL AVLCLLVVVA VAIGWVCRDR CLQHSYAGAW AVSPETELTE
       309
         |
SWNLLLLLS
```

- • in FASTA format
- isoform sequences in Fasta format
- Go to the bottom of the page to submit this sequence to a variety of analysis tools

BLAST — BLAST submission on ExPASy/SIB or at NCBI (USA)

Sequence analysis tools: ProtParam, ProtScale, Compute pI/Mw, PeptideMass, PeptideCutter, Dotlet (Java)

ScanProsite, MotifScan

Direct Submission to SWISS-MODEL

ExPASy Home page | Site Map | Search ExPASy | Contact us | Proteomics tools | Swiss-Prot

Hosted by SIB Switzerland | Mirror sites: Australia | Bolivia | Canada | China | Korea | Taiwan | USA

Swiss-Prot & TrEMBL
introduce a new arithmetical concept !

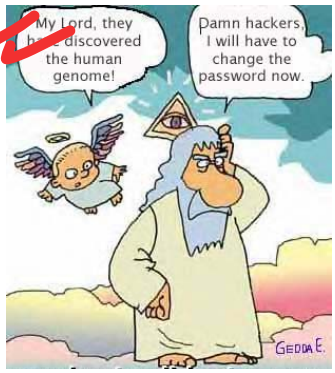
170'000 + 1'600'000 ≈ 1'200'000

Redundancy in TrEMBL
&
Redundancy between TrEMBL and Swiss-Prot

- In 2 years….more than 2'000'000 protein sequences
- But, in the future: redundancy is going to decrease:
  « new » genome sequencing -> « new » proteins
  (AB, sept 2002)

In the case of human proteins, the redundancy is still very high:

11'900 + 45'000 ≈ about 22'000*

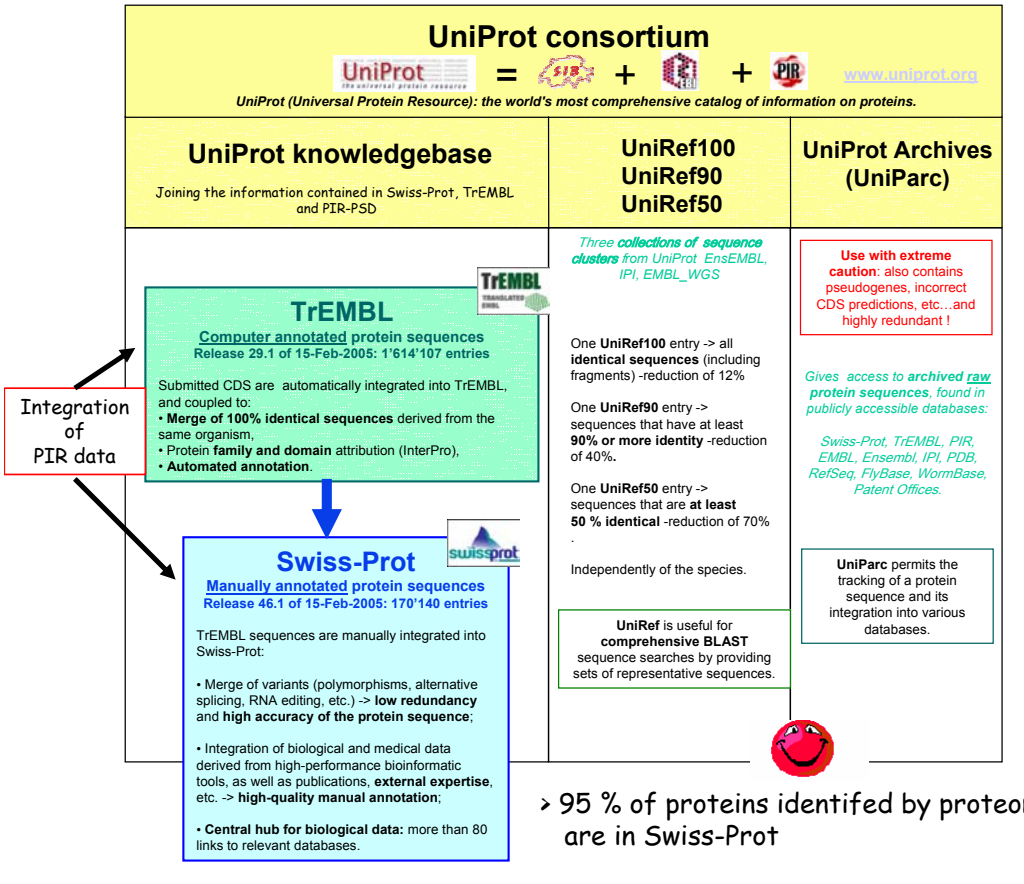


* human gene number estimation:
25'000-35'000

MS proteomics has verified more than 10% of human genes products, but has not identified significant numbers of unpredicted proteins (Southan C, Proteomics, 2004)

Are missing:

- Sequences not submitted to EMBL/GenBank/DDJB (and PIR)
- Not yet predicted or known genes (« no CDS provided by the submitters» or no DNA sequence)
- Confidential data (Patent application sequences)
- Immunoglobulins, T-cell receptors (-> UniParc)
- …



› 95 % of proteins identifed by proteomic studies are in Swiss-Prot

Take home message

- Be aware of the differences between TrEMBL and Swiss-Prot.
- Always cite the Accession number, not the ID.
- We need **your** feedback!

swiss-prot@expasy.org

Righting the wrongs

"Sequences are rarely deposited in a "mature" state; as with all scientific research, DNA and protein annotation is a continual process of learning, revision and corrections."

"Sequencing error rates: ~1 base in 10'000"

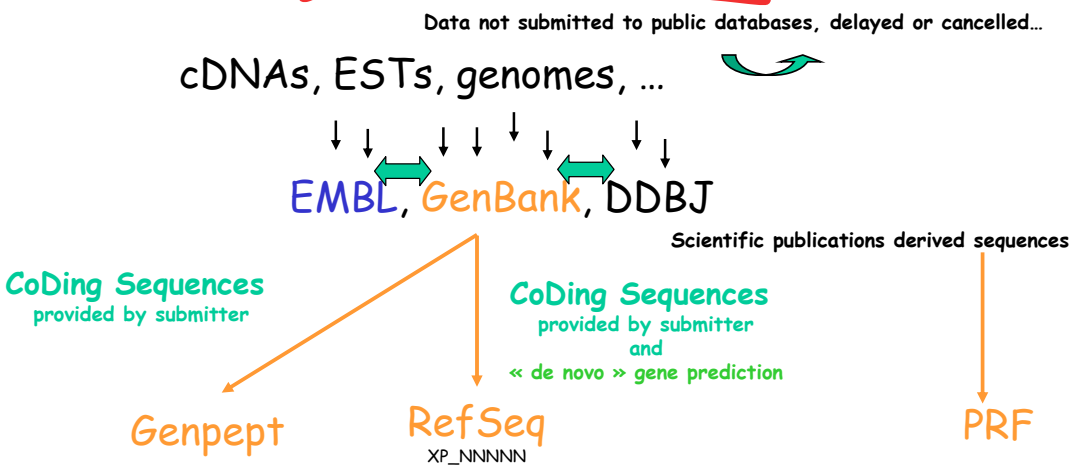
"Making people aware of errors is good and great; making people aware that they're responsible also for correcting errors is even greater"

C. Hardley, EMBO reports, 4(9), 2003.

Protein sequence databases

The NCBI-nr pathway (Entrez protein)

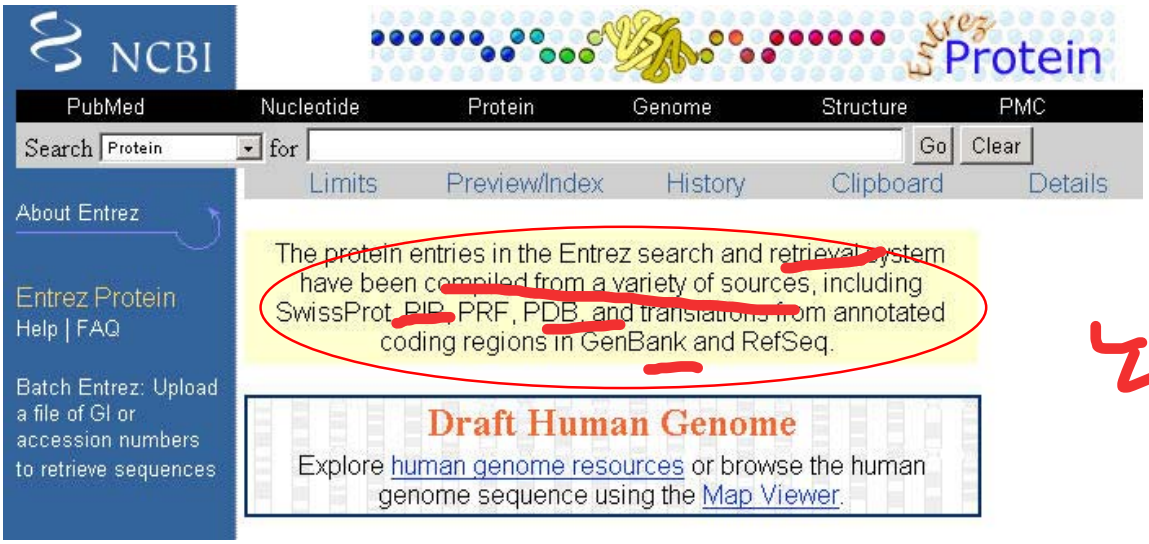
Real life of a protein sequence …



NCBI nr: Swiss-Prot + GenPept + (PIR) + RefSeq + PDB + PRF

Protein sequences: « NR database »
Entrez protein

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein



Scientific publications
derived sequences
« Journal scan »
(integrated into TrEMBL)

NCBI-nr: Swiss-Prot + GenPept + (PIR) + RefSeq + PDB + PRF

All PIR data have been integrated into Swiss-Prot and TrEMBL (UniProt)

derived from GenBank/EMBL/DDBJ sequences which have a CDS annotated on them - equivalent to TrEMBL, except that it is redundant with Swiss-Prot

3D structure database: all the protein sequences which have been cristallized (Swiss-Prot/TrEMBL are crosslinked to PDB)

RefSeq/Protein: http://www.ncbi.nlm.nih.gov/RefSeq/

- The RefSeq collection, which is tighly linked to LocusLink contains: genomic DNA, transcript (RNA), and protein products

- RefSeq provides a non-redundant set of sequences, derived from GenBank, the literature and gene prediction

- Release 3 includes over 800'000 proteins from 2218 organisms (including 1100 viruses and 150 bacteria).

AC
KW
Taxonomy

References

GenBank source

1: NP_057603. ameloblastin [Hom...[gi:7706067]

```
LOCUS       NP_057603               447 aa            linear   PRI 05-OCT-2003
DEFINITION  ameloblastin [Homo sapiens].
ACCESSION   NP_057603
VERSION     NP_057603.1  GI:7706067
DBSOURCE    REFSEQ: accession NM_016519.2
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 447)
  AUTHORS   Toyosawa,S., Fujiwara,T., Ooshima,T., Shintani,S., Sato,A.,
            Ogawa,Y., Sobue,S. and Ijuhin,N.
  TITLE     Cloning and characterization of the human ameloblastin gene
  JOURNAL   Gene 256 (1-2), 1-11 (2000)
  MEDLINE   20509985
   PUBMED   11054529
REFERENCE   2  (residues 1 to 447)
  AUTHORS   MacDougall,M., Simmons,D., Gu,T.T., Forsman-Semb,K., Mardh,C.K.,
            Mesbah,M., Forest,N., Krebsbach,P.H., Yamada,Y. and Berdal,A.
  TITLE     Cloning, characterization and immunolocalization of human
            ameloblastin
  JOURNAL   Eur. J. Oral Sci. 108 (4), 303-310 (2000)
  MEDLINE   20401566
   PUBMED   10946765
REFERENCE   3  (residues 1 to 447)
  AUTHORS   Bartlett,J.D. and Simmer,J.P.
  TITLE     Proteinases in developing dental enamel
  JOURNAL   Crit. Rev. Oral Biol. Med. 10 (4), 425-441 (1999)
  MEDLINE   20098047
   PUBMED   10634581
REFERENCE   4  (residues 1 to 447)
  AUTHORS   MacDougall,M., DuPont,B.R., Simmons,D., Reus,B., Krebsbach,P.,
            Karrman,C., Holmgren,G., Leach,R.J. and Forsman,K.
  TITLE     Ameloblastin gene (AMBN) maps within the critical region for
            autosomal dominant amelogenesis imperfecta at chromosome 4q21
  JOURNAL   Genomics 41 (1), 115-118 (1997)
  MEDLINE   97271565
   PUBMED   9126491
REFERENCE   5  (residues 1 to 447)
  AUTHORS   Lee,S.K., Krebsbach,P.H., Matsuki,Y., Nanci,A., Yamada,K.M. and
            Yamada,Y.
  TITLE     Ameloblastin expression in rat incisors and human tooth germs
  JOURNAL   Int. J. Dev. Biol. 40 (6), 1141-1150 (1996)
  MEDLINE   97184059
   PUBMED   9032019
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AF219994.2 and AF263464.1.
```

RefSeq/Protein

GenBank source

As for the nucleic acid sequence, RefSeq chooses a protein Reference Sequence: they do not annotate the sequence differences.

- If there is an alternative splicing event, there will be several entries for a same gene

Related Sequences ?

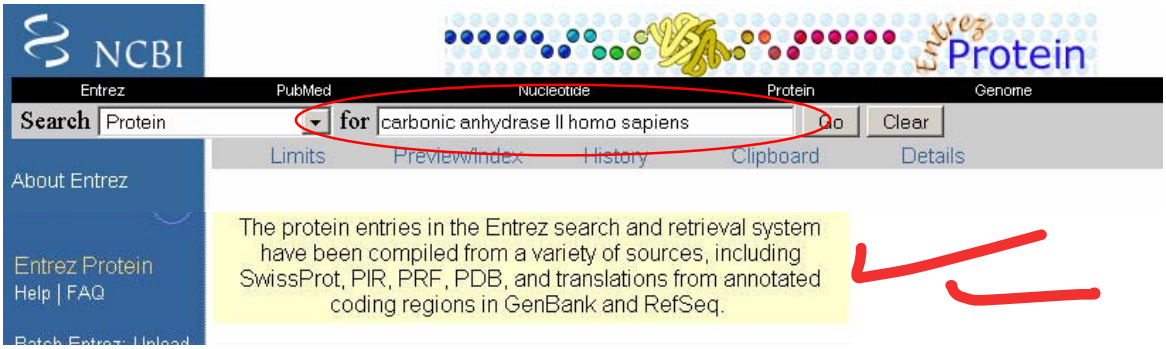
Nucleotide	Type	Protein	
AY009124	g	AAG27036	BL
AF209780	m	AAG35772	BL
→ AF219994	m	AAF37355	BL
→ AF263464	m	AAF73048	BL
AJ537437	m	CAD60938	BL
None	p	Q9NP70	BL

Annotation

```
                Summary: Ameloblastin is thought to represent an unique
                ameloblast-specific gene product that may be important in enamel
                matrix formation and mineralization. The gene is located on
                chromosome 4 near other genes associated with mineralized tissues:
                osteopontin, bone sialoprotein, and bone morphogenetic protein 3.
                Based on its cytogenetic location, this gene is a candidate gene
                for one form of the disorder, dentinogenesis imperfecta, and/or the
                disorder, autosomal dominant amylogenesis imperfecta.
FEATURES             Location/Qualifiers
     source          1..447
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="4"
                     /map="4q21"
     Protein         1..447
                     /product="ameloblastin"
     variation       255
                     /allele="V"
                     /allele="A"
                     /db_xref="dbSNP:7439186"
     CDS             1..447
                     /gene="AMBN"
                     /coded_by="NM_016519.2:76..1419"
                     /note="go_component: extracellular matrix [goid 0005578]
                     [evidence NAS];
                     go_function: structural constituent of tooth enamel [goid
                     0030345] [evidence NAS];
                     go_process: bone mineralization [goid 0030282] [evidence
                     NAS]"
Cross references     /db_xref="GeneID:258"
                     /db_xref="LocusID:258"
                     /db_xref="MIM:601259"
ORIGIN
        1 msaskiplfk mkdlililcl lemsfavpff pqqsgtpgma slsletmrql gslqrlntls
       61 qysrygfgks fnslwmhgll pphsslpwmr prehetqqye yslpvhpppl psqpslkpqq
      121 pglkpflqsa aattnqatal kealqppihl ghlplqegel plvqqqvaps dkppkpelpg
      181 vdfadpqgps lpgmdfpdpq gpslpgldfa dpqgstifqi arlishgpmp qnkqsplypg
      241 mlyvpfganq lnaparlgim sseevaggre dpmaygamfp gfggmrpgfe gmphnpamgg
      301 dftlefdspv aatkgpenee ggaqgspmpe anpdnlenpa fltelepaph agllalpkdd
      361 ipglprspsg kmkglpsvtp aaadplmtpe ladvyrtyda dmttsvdfqe eatmdttmap
      421 nslqtsmpgn kaqepemmhd awhfqep
//
```

Query at Entrez protein

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein



Typical result of a query at « Entrez protein »

RefSeq

Swiss-Prot

Genpept (gb/embl/ddbj)

PDB

☐ **3:** NP_000058
carbonic anhydrase II; carbonate dehydratase II; carbonic dehydratase; carbonic anhydrase B [Homo sapiens]
gi|4557395|ref|NP_000058.1|[4557395]

☐ **4:** NP_000333
solute carrier family 4, anion exchanger, member 1 (erythrocyte membrane protein band 3, Diego blood group); [Homo sapiens]
gi|4507021|ref|NP_000333.1|[4507021]

☐ **5:** NP_002334
lactotransferrin; lactoferrin [Homo sapiens]
gi|4505043|ref|NP_002334.1|[4505043]

☐ **6:** NP_003606
solute carrier family 4, sodium bicarbonate cotransporter, member 7; sodium bicarbonate cotransporter 2; solute member 6 [Homo sapiens]
gi|19923176|ref|NP_003606.2|[19923176]

☐ **7:** P00918
Carbonic anhydrase II (Carbonate dehydratase II) (CA-II) (Carbonic anhydrase C)
gi|115456|sp|P00918|CAH2_HUMAN[115456]

☐ **8:** AAH11949
Carbonic anhydrase II [Homo sapiens]
gi|15080386|gb|AAH11949.1|[15080386]

☐ **9:** 1OQ5A
Chain A, Carbonic Anhydrase Ii In Complex With Nanomolar Inhibitor
gi|47168469|pdb|1OQ5|A[47168469]

The AC number jungle

Type of record	Sample Accession Format
GenBank/EMBL/DDBJ	One letter followed by five digits: e.g. U12345 Two letters followed by 6 digits: e.g. AF123456
Swiss-Prot/TrEMBL	One letter and five digits/letters: e.g. P12345
RefSeq nucleotide	Two letters, underscore bar and six digit: e.g. mRNA NM_000492 e.g. genomic NT_000907
RefSeq protein	e.g. NP_00483
RefSeq prediction	e.g. XM_000483 e.g. XP_000467
PDB (protein structure)	One digit followed by three letters: e.g. 1TUP

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Databases 2: 'genomics'

- Contain informations on gene chromosomal location (mapping) and nomenclature, and provide links to sequence databases; *has usually no sequence*;
- Exist for most organisms important in life science research; usually species specific.
- Examples: MIM, GDB (human), MGD (mouse), FlyBase (Drosophila), SGD (yeast), MaizeDB (maize), SubtiList (B.subtilis), etc.;
- Generally relational db (Oracle, SyBase or AceDb).

MIM / OMIM

- OMIM™: Online Mendelian Inheritance in Man
- catalog of human genes and genetic disorders
- contains a summary of literature and reference information. It also contains links to publications and sequence information.

*133[illegible] ERYTHROPOI[illegible] EPO

TABLE OF CONTENTS

Database Links

MEDLINE | Protein | DNA | Genome | LocusLink | Gene Map | GDB | MGD | Nomenclature

Gene Map Locus: 7q21

Note: pressing the symbol will find the citations in MEDLINE whose text most closely matches the text of the preceding OMIM paragraph, using the Entrez MEDLINE neighboring function.

TEXT

DESCRIPTION

Human erythropoietin is an acidic glycoprotein hormone with a molecular mass of 34 kD. As the prime regulator of red cell production, its major functions are to promote erythroid differentiation and to initiate hemoglobin synthesis.

CLONING

Lee-Huang (1984) cloned human erythropoietin cDNA in E. coli. McDonald et al. (1986) and Shoemaker and Mitsock (1986) cloned the mouse gene and the latter workers showed that coding DNA and amino acid sequence are about 80% conserved between man and mouse. This is a much higher order of conservation than for various interferons, interleukin-2, and GM-CSF. Sherwood and Shoveral (1986) described a human renal carcinoma cell line that continuously produces erythropoietin. Eschbach et al. (1987) demonstrated the effectiveness of recombinant human erythropoietin in treating the anemia of end-stage renal disease.

Romanowski and Sytkowski (1994) reviewed the molecular structure of human erythropoietin in historical perspective. The EPO gene has 5 exons that code for a 193-amino acid propolypeptide. A 27-amino acid leader sequence is cleaved off the amino terminus of the propeptide, yielding the functional 166-amino acid protein. However, recombinant human EPO expressed in Chinese hamster ovary cells contains only 165 amino acids, having lost arg166. The mechanism for this was undefined, and whether EPO circulating in the plasma also lacked arg166 was not known. Both the nucleotide and amino acid sequences of EPO are highly conserved among mammals.

In the central nervous system, neurons express EPO receptor (EPOR) and astrocytes produce EPO. EPO has been shown to protect primary cultured neurons from NMDA receptor-mediated glutamate toxicity. Sakanaka et al. (1998) reported in vivo evidence that EPO protects neurons against ischemia-induced cell death. They presented findings suggesting that EPO may exert its neuroprotective effect by reducing the nitric oxide-mediated formation of free radicals or antagonizing their toxicity. Siren et al. (2001) presented data suggesting that inhibition of neuronal apoptosis underlies short latency protective effects of EPO after cerebral ischemia and other brain injuries. They suggested that evaluation of EPO, a compound established as clinically safe, as neuroprotective therapy in acute brain injury is indicated.

Novel erythropoiesis stimulating protein (NESP) stimulates erythropoiesis in the same manner as human recombinant EPO. NESP is distinct from EPO in that it has additional sialic acid which has been shown to confer an increased terminal half-life in animal models, patients with chronic renal failure, and cancer patients receiving multiple cycles of chemotherapy (Macdougall et al., 1999). In studies of 89 patients with nonmyeloid malignancies, Smith et al. (2001) found that NESP was well tolerated, with response rates ranging from 61 to 83%, depending on dosage.

MAPPING

Law et al. (1986) assigned EPO to chromosome 7 by Southern blot analysis of DNA from human/Chinese hamster cell hybrids with a cDNA clone for the entire coding region of the gene. Further localization to 7q11-q22 was achieved by in situ hybridization. They found a RFLP with a frequency of about 20% in a Chinese population. By hybridization analysis (dot-blot) of DNA from human chromosomes isolated by high resolution dual laser sorting, Fowell et al. (1986) also located EPO on chromosome 7. By somatic cell hybrid analysis, Watkins et al. (1986) placed EPO on the proximal half of 7q, closely linked to COL1A2 (120160) and to DNA markers linked to CF (219700). Because of the close linkage of EPO to COL1A2 and markers linked to CF, it is probably justified to narrow the assignment of EPO to 7q21-q22.

By in situ hybridization and by genetic analysis using RFLPs in interspecific mouse backcross DNAs, Lacombe et al. (1988) demonstrated that EPO is located on chromosome 5 in the mouse.

In addition to its role as a kidney cytokine regulating hematopoiesis, EPO is also produced in the brain after oxidative or nitrosative stress. The transcription factor HIF1 (603348) upregulates EPO following hypoxic stimuli. Digicaylioglu and Lipton (2001) demonstrated that preconditioning with EPO protects neurons in models of ischemic and degenerative damage due to excitotoxins and consequent generation of free radicals, including nitric oxide. Activation of neuronal EPO receptors (133171) prevents apoptosis induced by NMDA or nitric oxide by triggering crosstalk between the signaling pathways JAK2 (147796) and NFKB (see 164011). Digicaylioglu and Lipton (2001) demonstrated that EPO receptor-mediated activation of JAK2 leads to phosphorylation of the inhibitor of NFKB (see I-kappa-B-alpha, 164008), subsequent nuclear translocation of the transcription factor NFKB, and NFKB-dependent transcription of neuroprotective genes. Transfection of cerebrocortical neurons with a dominant interfering form of JAK2 or an I-kappa-B-alpha superrepressor blocks EPO-mediated prevention of neuronal apoptosis. Thus, neuronal EPO receptors activate a neuroprotective pathway that is distinct from previously well characterized JAK and NFKB functions. Moreover, this EPO effect may underlie neuroprotection mediated by hypoxic-ischemic preconditioning.

ANIMAL MODEL

Synthesis of erythropoietin in the kidney and liver in response to hypoxia depends on both protein synthesis and heme synthesis. Goldwasser et al. (1988) proposed a model in which a ligand-dependent conformational change in a heme protein accounts for the mechanism by which hypoxia as well as cobalt and nickel stimulates the production of erythropoietin. Semenza et al. (1989) generated transgenic mice containing the human erythropoietin gene and found increased erythropoietin mRNA expression not only in liver and kidney but in all other transgenic tissues analyzed. The mice were polycythemic, with increased erythroid precursors in hematopoietic tissues and increased erythrocytic indices in peripheral blood. From further studies in these transgenic mice, Semenza et al. (1989) concluded that different DNA sequences flanking the EPO gene control liver versus kidney expression of the gene and that some of these sequences are located 3-prime to the gene.

Naffakh et al. (1995) examined whether the secretion of erythropoietin from genetically modified cells could represent an alternative to repeated injections of the recombinant hormone for treating chronic anemias responsive to EPO. Primary mouse skin fibroblasts were transduced with a retroviral vector in which the murine cDNA was expressed under the control of the murine phosphoglycerate kinase promoter. 'Neo-organs' containing the genetically modified fibroblasts embedded into collagen lattices were implanted into the peritoneal cavity of mice. Increased hematocrit and elevated serum EPO concentration were observed in recipient animals over a 10-month observation period. The approach was considered applicable to the treatment of human anemias.

Osborne et al. (1995) investigated in rats the expression and biologic effects of transplanting autologous vascular smooth muscle cells transduced with a retroviral vector encoding rat erythropoietin cDNA. Vector-derived Epo secretion caused increases in reticulocytes followed by clinically significant increases in hematocrit and hemoglobin for up to 11 weeks. There were no significant differences between control and treated animals in the number of white blood cells and platelets. Kidney and to a lesser extent liver are specific organs that synthesize Epo in response to tissue oxygenation. In the treated animals, endogenous Epo mRNA was largely downregulated in kidney and absent from liver. These results indicated to the authors that vascular smooth muscle cells can be genetically modified to provide treatment of anemias due to Epo deficiency and suggest that this cell type may be targeted in the treatment of other diseases requiring systemic therapeutic protein delivery.

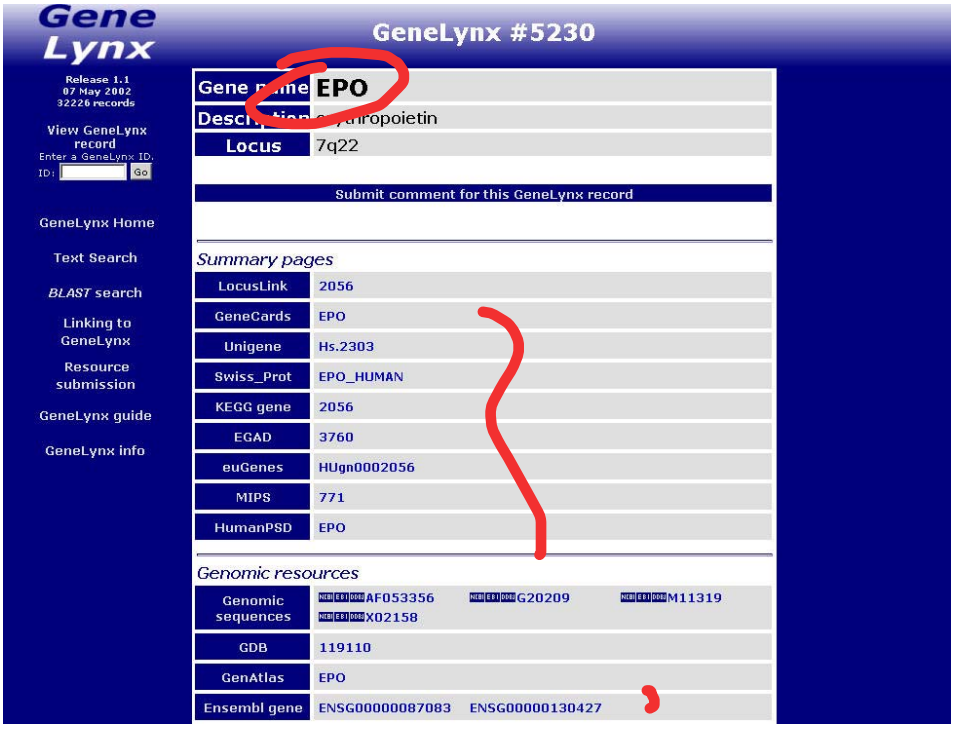
Similar experiments were performed by Kessler et al. (1996), who demonstrated that, following a single intramuscular administration of a recombinant adeno-associated virus (rAAV) vector containing the beta-galactosidase gene into adult mice, protein expression was detected in myofibers for at least 32 weeks. Furthermore, a single intramuscular administration of an AAV vector containing a gene for human erythropoietin into mice resulted in dose-dependent secretion of erythropoietin and corresponding increases in red blood cell production that persisted for up to 40 weeks. Primary human myocytes transduced in vitro with the AAV-Epo vector also showed dose-dependent production of Epo.

SEE ALSO

Jacobs et al. (1985); Lin et al. (1985); Semenza et al. (1989)

REFERENCES

http://www.genelynx.org/



Collections of hyperlinks for each human gene

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Mutation/polymorphism: definitions

- **SNPs**: single nucleotide polymorphisms: occur approximately once every 100 to 300 bases

(distinction between sequencing error and polymorphism !)

- **c-SNPs**: coding single nucleotide polymorphisms
  (Single Nucleotide Polymorphisms within cDNA sequences)
- **SAPs**: single amino-acid polymorphisms

- Missense mutation: -> SAP
- Nonsense mutation: -> STOP
- Insertion/deletion of nucleotides -> frameshift…

Databases 3: mutation/polymorphism

- Contain informations on sequence variations linked or not to genetic diseases;
- Mainly human but: OMIA - Online Mendelian Inheritance in Animals
- **General db**:
  - OMIM
  - HMGD - Human Gene Mutation db
  - SVD - Sequence variation db
  - HGBASE - Human Genic Bi-Allelic Sequences db
  - dbSNP - Human single nucleotide polymorphism (SNP) db
- **Disease-specific db**: most of these databases are either linked to a single gene or to a single disease;
  - p53 mutation db
  - ADB - Albinism db (Mutations in human genes causing albinism)
  - Asthma and Allergy gene db
  - ....

For human (Amos'link)

- HGMD - Human Gene Mutation db
- SVD - EBI Sequence variation db
- HGBASE - Human Genic Bi-Allelic Sequences db
- The SNP consortium
- dbSNP - Human single nucleotide polymorphism (SNP) db

- ALFRED - Allele Frequency Db
- SeattleSNPs - UW-FHCRC Variation Discovery Resource
- PicSNP - Catalog of non-synonymous SNP

- List of mutation databases from OMIM
- List of mutation databases from IMT (Finland)

Organism	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (ss#'s) with frequency	Number of (ss#'s) with genotype
Homo sapiens / *geneReport*	21,581,724	10,054,521 (5,054,675)	488,391	2,727,888
Gallus gallus / *geneReport*	3,641,959	3,291,672 (3,281,479)		3,624,831
Oryza sativa	3,931,108	3,899,916 (22,057)		
Anopheles gambiae	1,368,805	1,136,268 (0)		
Mus musculus / *geneReport*	600,576	581,577 (535,237)		545,500
Canis familiaris	975,417	957,828 (12,315)	17	
Rattus norvegicus	45,360	43,229 (702)		669
Saccharum hybrid cultivar	42,853	42,853 (0)		
Danio rerio	2,031	2,025 (1,903)		2,031
Sus scrofa	5,450	1,521 (24)		
Bos indicus x Bos taurus	2,425	2,425 (0)		
Bos taurus	2,058	2,058 (45)	54	
Caenorhabditis elegans	1,065	1,065 (0)		

Mutation/polymorphism

- No single source for all SNPs (~100 SNPs db) !
- Generally modest size; lack of coordination and format standards in these databases making it difficult to access the data.
- ! Numbering of the mutated amino acid depends on the db (aa no 1 is not necessary the initiator Met !)
- There are initiatives to unify these databases (politic/founding problems) Mutation Database Initiative (4th July 1996).

-> SVD - Sequence Variation Database project at EBI (HMutDB)
http://www.ebi.ac.uk/mutations/central/

-> HUGO Mutation Database Initiative (MDI).
Human Genome Variation Society
http://www.genomic.unimelb.edu.au/mdi/dblist/dblist.html

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Protein domain/family: some definitions

- Most proteins have « modular » structures
- Estimation: ~ 3 domains / protein

Some statistics

Proteome Analysis @EBI

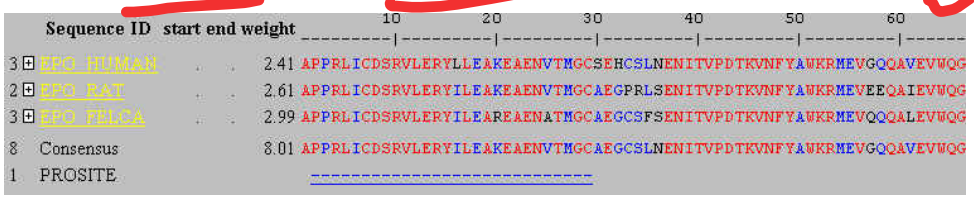
15 most common domains for *H. sapiens* [help]

InterPro	Matches per genome (Proteins matched)	Name
IPR007110	6473(983)	Immunoglobulin-like
IPR007087	29392(949)	Zn-finger, C2H2 type
IPR001841	1182(391)	Zn-finger, RING
IPR008938	879(391)	ARM repeat fold
IPR001849	1078(354)	Pleckstrin-like
IPR009057	406(342)	Homeodomain-like
IPR006209	3019(322)	EGF-like domain
IPR000504	1540(296)	RNA-binding region RNP-1 (RNA recognition motif)
IPR008941	690(291)	TPR-like
IPR001452	2206(290)	SH3
IPR008957	934(258)	Fibronectin, type III-like fold
IPR009058	216(209)	Winged helix DNA-binding
IPR001478	1335(208)	PDZ/DHR/GLGF domain
IPR008973	318(200)	C2 calcium/lipid-binding domain, CaLB
IPR000210	495(196)	BTB/POZ domain

http://www.ebi.ac.uk/proteome/HUMAN/interpro/top15d.html

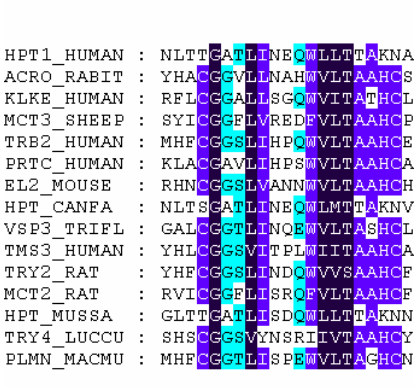
Protein domain/family: some definitions

- Domains (conserved sequences or structures) are identified by multiple sequence alignments



- Domains can be defined by different methods:
  - Pattern (regular expression); used for very conserved domains
  - Profiles (weighted matrices); two dimensional tables of position specific match-, gap-, and insertion-scores, derived from aligned sequence families; used for less conserved domains
  - Hidden Markov Model (HMM); probabilistic models; an other method to generate profiles.

Pattern-Profile



- Pattern **[LIVM]-[ST]-A-[STAG]-H-C**

  ⟶ Yes or no

- Profile:

```
ID   TRYPSIN_DOM; MATRIX.
AC   PS50240;
DT   DEC-2001 (CREATED); DEC-2001 (DATA UPDATE); DEC-2001 (INFO UPDATE).
DE   Serine proteases, trypsin domain profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=234;
MA   /DISJOINT: DEFINITION=PROTECT; N1=6; N2=229;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.0169; R2=0.00836256; TEXT='-LogE';
MA   /CUT_OFF: LEVEL=0; SCORE=1134; N_SCORE=9.5; MODE=1; TEXT='!';
MA   /CUT_OFF: LEVEL=-1; SCORE=775; N_SCORE=6.5; MODE=1; TEXT='?';
MA   /DEFAULT: M0=-9; D=-20; I=-20; B1=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105;
MA   /I:       B1=0; BI=-105; BD=-105;
MA             A  B  D  E  F  G  H  I  K  L  M  N  P  Q  R  S  T  V  W  Y
MA   /M: SY='I'; M= -8,-29,-34,-26,  3,-34,-24, 34,-26, 19, 15,-24,-21,-21,-24,-19, -8, 25,-19,  3;
MA   /M: SY='N'; M=  0, 14, 10,  1,-22, -1,  6,-23, -4,-26,-17, 20,-14, -1, -6, 13,  2,-20,-34,-15;
MA   /M: SY='E'; M= -4,  4,  7, 14,-26,-13, -7,-23,  3,-22,-16,  2,  7,  3, -3,  2, -2,-21,-30,-18;
MA   /M: SY='R'; M=-12,  5,  5,  7,-23,-17,  3,-24,  8,-20,-12,  7,-16, 10, 12, -2, -6,-21,-27, -9;
MA   /M: SY='W'; M=-16,-33,-35,-27, 13,-22,-24,-11,-18,-13,-13,-31,-27,-20,-18,-30,-21,-18, 97, 25;
MA   /M: SY='V'; M=  1,-29,-31,-28, -1,-30,-29, 31,-22, 13, 11,-27,-27,-26,-22,-12, -2, 41,-27, -8;
MA   /M: SY='L'; M= -8,-29,-31,-22,  9,-30,-21, 23,-27, 37, 20,-28,-28,-21,-20,-25, -8, 17,-20, -1;
MA   /M: SY='T'; M=  2, -1, -9, -9,-11,-17,-19,-10,-10,-13,-11,  1,-11, -9,-10, 23, 43,  0,-32,-12;
MA   /M: SY='A'; M= 45, -9,-19,-10,-20, -2,-15,-11,-10,-11,-10, -9,-11, -9,-19, 10,  1, -1,-21,-18;
MA   /M: SY='A'; M= 40, -9,-17, -8,-21,  5,-18,-14, -9,-13,-12, -8,-11, -9,-16,  9, -2, -5,-21,-21;
MA   /M: SY='H'; M=-18,  0,  0,  1,-21,-19, 89,-29, -8,-21, -1,  9,-19, 11,  0, -7,-17,-29,-30, 16;
MA   /M: SY='C'; M= -9,-18,-28,-29,-20,-29,-29,-29,-29,-20,-19,-18,-39,-29,-29,  -9,  -9,  -9,-49,-29;
MA   /I:       E1=0; IE=-105; DE=-105;
//
```

⟶ score/threshold

Protein domain/family databases

- Contains biologically significant « pattern / profiles/ HMM » formulated in such a way that, with appropriate computional tools, it can rapidly and reliably determine to which known family of proteins (if any) a new sequence belongs to
- Used as a tool to identify the function of uncharacterized proteins translated from genomic or cDNA sequences (« functional diagnostic »)
- Either manually curated (i.e. PROSITE, PfamA, PRINTS, SMART, TIGRFAM etc.) or automatically generated (i.e. PfamB, ProDom, DOMO)

Protein domain/family db

PROSITE	Patterns / Profiles
ProDom	Aligned motifs (PSI-BLAST) (Pfam B)
PRINTS	Aligned motifs
Pfam	HMM (Hidden Markov Models)
SMART	HMM
TIGRfam	HMM

Interpro

DOMO	Aligned motifs
BLOCKS	Aligned motifs (PSI-BLAST)
CDD	Pfam and SMART

-> A Conserved Domain Database and Search Service

Prosite http://www.expasy.org/prosite/

- Created in 1988 (SIB)
- Contains functional domains fully annotated, based on two methods: patterns and profiles
- Entries are deposited in PROSITE in two distinct files:
  - Pattern/profiles with the list of all matches in SWISS-PROT
  - Documentation

15-Aug-2004: contains 1277 documentation entries that describe 1736 different patterns, rules and profiles/matrices.

NiceSite View of PROSITE: PS00817

General information about the entry	
Entry name	EPO_TPO
Accession number	PS00817
Entry type	PATTERN
Date	OCT-1993 (CREATED); NOV-1995 (DATA UPDATE); JUL-1998 (INFO UPDATE).
PROSITE documentation	PDOC00644
Name and characterization of the entry	
Description	Erythropoietin / thrombopoietin signature.
Pattern	P-x(4)-C-D-x-R-[LIVM](2)-x-[KR]-x(14)-C.
Numerical results	

- SWISS-PROT release number: **38**; total number of sequence entries in that release: **80000**.
- Total number of hits in SWISS-PROT: **14 hits in 14 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **14 hits in 14 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **0**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **1**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **100.00 %**

Diagnostic performance

Comments

- Taxonomic range: **Eukaryotes**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- 'Interesting' site in the pattern: **3,disulfide**
- 'Interesting' site in the pattern: **11,disulfide**

List of matches

Cross-references

SWISS-PROT

True positive hits:

EPO_BOVIN (P48617), EPO_CANFA (P33707), EPO_FELCA (P33708),
EPO_HUMAN (P01588), EPO_MACFA (P07865), EPO_MACMU (Q28513),
EPO_MOUSE (P07321), EPO_PIG (P49157), EPO_RAT (P29676),
EPO_SHEEP (P33709), TPO_CANFA (P42705), TPO_HUMAN (P40225),
TPO_MOUSE (P40226), TPO_RAT (P49745)

'Potential' hits (sequences that belong to the set under consideration, but which were not picked up because the region(s) that are used as a 'fingerprint' (pattern or profile) is not yet available in the data bank (partial sequence)):

TPO_PIG (P42706)

NiceSite View of PROSITE: PS50097

General information about the entry	
Entry name	BTB
Accession number	PS50097
Entry type	MATRIX
Date	SEP-2000 (CREATED); SEP-2000 (DATA UPDATE); SEP-2000 (INFO UPDATE).
PROSITE documentation	PDOC50097
Name and characterization of the entry	
Description	BTB domain profile.
Matrix / Profile	/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWY2', LENGTH=67, /DISJOINT: DEFINITION=PROTECT, N1=6, N2=62; /NORMALIZATION: MODE=1, FUNCTION=LINEAR, R1= 9751, R2= 02068202, TEXT='-LogE', /CUT_OFF: LEVEL=0, SCORE=363, N_SCORE=8 5, MODE=1, TEXT='!'; /CUT_OFF: LEVEL=-1; SCORE=267; N_SCORE=6 5; MODE=1; TEXT='?'; /DEFAULT: D=-20; I=-20; B1=-50; E1=-50; MI=-105; MD=-105; DM=-105; DMI=-105; MM=1; M0=-2; /I: B1=0; BI=-105; BD=-105; /M: SY='C'; M=-6,-10,28,-14,-9,-15,-20,-14,-19,-15,-17,-14,-8,-19,-14,-15,0,0,-9,-32,-17,-12; /M: SY='D'; M=-16,41,-28,53,15,-34,-11,-1,-33,0,-27,-25,21,-11,0,-8,2,-6,-26,-38,-19,7; /M: SY='V'; M=2,-23,-8,-28,-24,-1,-24,-25,16,-20,7,6,-20,-25,-23,-20,-10,-4,24,-23,-9,-24; /M: SY='T'; M=-2,-13,-18,-19,-13,-7,-24,-19,6,-8,-2,1,-11,-17,-11,-10,-1,10,10,-24,-6,-13; /M: SY='L'; M=-11,-30,-22,-33,-24,15,-32,-23,25,-29,35,17,-26,-27,-23,-22,-24,-9,16,-17,3,-24; /M: SY='V'; M=0,-11,-18,-13,-10,-12,-20,-13,1,-6,-4,2,-10,-19,-6,-7,-4,-2,8,-25,-9,-9; /M: SY='V'; M=1,-25,-3,-29,-25,-2,-26,-26,17,-22,10,7,-23,-25,-23,-22,-11,-3,24,-27,-10,-25; /M: SY='D'; M=-6,7,-26,8,7,-25,6,-7,-27,0,-23,-17,8,-13,0,-3,3,-6,-23,-27,-17,3; /I: I=-5; MI=0; IM=0; DM=-15; MD=-15; /M: SY='G'; M=-6,8,-27,8,-3,-27,22,-7,-30,-8,-26,-19,10,-14,-8,-9,2,-9,-24,-28,-21,-6; /M: SY='K'; M=-7,-4,-23,-4,7,-23,-13,-2,-21,10,-18,-9,-3,-12,7,9,-2,-4,-16,-25,-12,6; /M: SY='E'; M=-8,-6,-21,-8,1,-15,-21,-7,-7,-1,-10,-5,3,-14,0,-1,-2,-2,-6,-26,-9,-1; /M: SY='F'; M=-12,-28,-22,-34,-26,31,-31,-21,18,-26,16,9,-22,-27,-27,-21,-20,-9,14,-6,13,-26; /M: SY='R'; M=-13,-9,-24,-10,-3,-11,21,7,-17,7,-16,-4,-4,-8,2,9,-9,-9,-16,-20,-1,-2; /M: SY='A'; M=21,-15,-8,-22,-17,-10,-10,-23,0,-15,-5,-5,-14,-18,-17,-19,4,6,12,-24,-15,-17; /M: SY='H'; M=-15,5,-22,2,-1,-20,-16,65,-26,-8,-21,-5,15,-19,6,-2,-2,-11,-26,-32,7,0; /M: SY='K'; M=-12,-5,-29,-5,5,-25,-18,-8,-26,34,-24,-9,-1,-14,8,34,-8,-8,-17,-20,-10,5; /M: SY='A'; M=4,-12,-12,-16,-10,-6,-18,-14,-2,-13,-1,-2,-11,-17,-12,-13,-3,1,2,-24,-8,-11; /M: SY='V'; M=-7,-26,-19,-31,-26,7,-32,-24,27,-23,14,11,-22,-25,-23,-22,-13,0,28,-19,3,-26; /M: SY='L'; M=-10,-30,-20,-30,-21,9,-30,-20,22,-29,47,20,-29,-29,-20,-20,-29,-10,12,-20,0,-21; /M: SY='A'; M=18,-6,0,-12,-8,-18,-6,-16,-15,-10,-18,-12,-2,-14,-8,-13,18,11,-5,-32,-19,-8; /I: I=-6; MI=-32; MD=-32; IM=-32; /M: SY='A'; M=15,-9,-13,-14,-8,-12,-10,-12,-9,-4,-7,-3,-7,-14,-7,-9,1,-1,-4,-20,-9,-7; D=-6; /I: DM=-32; /M: SY='C'; M=-7,-11,6,-16,-14,-5,-17,1,-17,-10,-12,-8,-5,-24,-11,-3,-5,-7,-10,-25,-4,-13; /M: SY='S'; M=7,3,-9,4,1,-22,-3,-9,-22,-10,-29,-20,9,-9,-1,-11,32,15,-12,-39,-20,0; /M: SY='P'; M=-6,-1,-28,2,8,-26,-9,-10,-22,-3,-24,-17,-2,20,-1,-8,3,-2,-22,-30,-21,2; /M: SY='Y'; M=-14,-24,-24,-26,-23,30,-29,1,6,-17,5,3,-22,-29,-19,-15,-18,-9,2,13,45,-23; /I: I=-10; MI=-15; MD=-29; IM=-15; /M: SY='F'; M=-18,-29,-19,-37,-28,66,-29,-19,4,-28,14,5,-21,-29,-35,-19,-20,-10,2,5,24,-27; D=-6; /I: DM=-29; /M: SY='R'; M=-10,-4,-28,-4,4,-17,-18,-1,-22,17,-18,-9,-2,-16,8,19,-7,-9,-18,-14,-2,4; /M: SY='A'; M=7,-1,-18,-5,-2,-18,-13,-12,-12,-1,-12,-9,-1,-13,-3,-2,5,5,-6,-27,-14,-3; /M: SY='L'; M=-8,-22,-22,-27,-17,2,-26,-12,15,-16,25,23,-21,-23,-11,-12,-20,-8,7,-20,-1,-15; /M: SY='F'; M=-15,-28,-21,-35,-25,44,-30,-20,9,-27,19,8,-23,-27,-29,-20,-22,-9,5,0,17,-25; /M: SY='S'; M=-4,-3,-19,-5,0,-15,-16,-8,-13,-1,-13,-8,1,-14,1,-1,3,3,-10,-26,-10,0; /M: SY='N'; M=-4,8,-20,7,1,-25,2,-2,-24,-5,-23,-15,11,-13,1,-7,9,1,-19,-32,-17,1; /M: SY='B'; M=-6,6,-23,5,6,-24,-8,2,-22,0,-20,-12,6,-9,5,-3,1,-4,-20,-28,-13,5; /M: SY='F'; M=-11,-11,-25,-13,-10,5,-15,-10,-10,-12,-7,-4,-7,-12,-11,-10,-8,-5,-11,-10,0,-11; /I: MD=-15; /M: M=-5,-5,-14,-7,-3,-18,-13,-10,-14,-2,-14,-7,-3,-11,-3,-4,-1,-1,-11,-27,-13,-4; D=-1; /I: MD=-15;

Numerical results	

- SWISS-PROT release number: **39**; total number of sequence entries in that release: **87397**.
- Total number of hits in SWISS-PROT: **46 hits in 44 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **45 hits in 43 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **1 hits in 1 different sequences**
- Number of false hits (on unrelated proteins): **0 hits in 0 different sequences**
- Number of known missed hits: **0**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **0**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **100.00 %**

Comments

- Taxonomic range: **Eukaryotes, Eukaryotic viruses**
- Maximum known number of repetitions of the pattern in a single protein: **2**
- FT_KEY: **DOMAIN**
- FT_DESC: **BTB**

Cross-references

SWISS-PROT

True positive hits:

BAC1_HUMAN (Q14857), BAC1_MOUSE (P97303), BAC2_MOUSE (P97305),
BCL6_HUMAN (P41182), BCL6_MOUSE (P41183), BBC1_DROME (Q01295),
BBC2_DROME (Q01296), BBC3_DROME (Q01297), CALI_BOVIN (Q28068),
CALI_HUMAN (Q13939), GAGA_DROME (Q08605), GCL1_DROME (Q01820),
HKR3_HUMAN (P10074), KELC_DROME (Q04652), LOL2_DROME (P42283),
LOL5_DROME (P42284), PI10_HUMAN (P14892), PLIP_HUMAN (Q05516),
SPOP_HUMAN (Q13791), TTKA_DROME (P42282), TTKB_DROME (P17789),
VAS5_VACCC (P21073), VAS5_VACCV (P14768), VCO2_VACCC (P21077),
VCO2_VACCV (P17371), VCD4_SPVEA (P32228), VC13_SPVEA (P32206),
VF03_VACCC (P21013), VF03_VACCV (P14357), VNT0_HTLVL (P22413),
VNTP_HTLVL (P08073), Y4A1_HUMAN (Q43167), YA2K_SCHPO (Q10025),
Y1A1_YEAST (P40560), YHV2_CAEEL (P34529), YL30_CAEEL (P34573),
YMV5_CAEEL (P34548), YPT9_CAEEL (P41884), YR47_CAEEL (Q09563),
YSW1_CAEEL (Q10017), Z151_HUMAN (Q13105), Z151_MOUSE (Q60811),
ZN44_HUMAN (P14378)

Sequences which could potentially belong to the set under consideration:

TNF1_HUMAN (Q13829)

PFAM (HMMs): an entry

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

Protein families database of alignments and HMMs

EPO_TPO

Home Search by Browse by ftp iPfam Help

Figure 1: 1eer
Complex (cytokine/receptor)
Crystal structure of human erythropoietin complexed to its receptor at 1.9 angstroms

1buy Display pdb

Accession number: PF00758

Erythropoietin/thrombopoietin

Add Annotation

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

INTERPRO description (entry IPR001323)

Erythropoietin, a plasma glycoprotein, is the primary physiological mediator of erythropoiesis PUBMED:3773894. It is involved in the regulation of the level of peripheral erythrocytes by stimulating the differentiation of erythroid progenitor cells, found in the spleen and bone marrow, into mature erythrocytes PUBMED:3346214. It is primarily produced in adult kidneys and foetal liver, acting by attachment to specific binding sites on erythroid progenitor cells, stimulating their differentiation PUBMED:2877922. Severe kidney dysfunction causes reduction in the plasma levels of erythropoietin, resulting in chronic anaemia - injection of purified erythropoietin into the blood stream can help to relieve this type of anaemia. Levels of erythropoietin in plasma fluctuate with varying oxygen tension of the blood, but androgens and prostaglandins also modulate the levels to some extent PUBMED:2877922. Erythropoietin glycoprotein sequences are well conserved, a consequence of which is that the hormones are cross-reactive among mammals, i.e. that from one species, say human, can stimulate erythropoiesis in other species, say mouse or rat PUBMED:1420369.

Thrombopoeitin (TPO), a glycoprotein, is the mammalian0 hormone which functions as a megakaryocytic lineage specific growth and differentiation factor affecting the proliferation and maturation from their committed progenitor cells acting at a late stage of megakaryocyte development. It acts as a circulating regulator of platelet numbers.

QuickGO	
FUNCTION :	hormone activity (GO:0005179)
COMPONENT :	extracellular (GO:0005576)

For additional annotation, see the PROSITE document PDOC00644 [Expasy | SRS-UK | SRS-USA]

Alignment	Domain organisation
Seed (7) Full (32)	View 1 representative architecture View architectures for 32 proteins
Format: Coloured alignment	Zoom: 0.5 pixels/aa.
Get alignment View HMM logo	View Graphic
Further alignment options here Help relating to Pfam alignments here	

HMM

```
HMMER2.0  [2.2f]
NAME  EPO_TPO
ACC   PF00758
DESC  Erythropoietin/thrombopoietin
LENG  180
ALPH  Amino
RF    no
CS    no
MAP   yes
COM   hmmbuild -F HMM.ann SEED.ann
COM   hmmcalibrate --seed 0 HMM.ann
NSEQ  7
DATE  Wed Aug  1 04:23:19 2001
CKSUM 74
GA    25.0 25.0
TC    106.1 106.1
NC    0.0 0.0
XT      -8455     -4  -1000  -1000  -8455     -4  -8455     -4
NULT       -4  -8455
NULE      595  -1558     85     338   -294     453  -1158     197     249     902  -1085   -142     -21   -31
EVD    -90.885963   0.176335
HMM         A      C      D      E      F      G      H      I      K      L      M      N      P
        m->m   m->i   m->d   i->m   i->i   d->m   d->d   b->m   m->e
            -415      *  -2000
      1   -304  -1254    633   1460  -1327  -1319    -38  -1020    242    452   -389    -60  -1434
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378   -415      *      *
      2  -2228  -1908  -3555  -3373   -604  -3255  -2515    -62  -2963   2976    358  -3207  -3275  -26
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
      3  -2072  -1684  -4075  -3524   -176  -3842  -2605    425  -3152   2687   2365  -3417  -3383  -24
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
      4  -2228  -1908  -3555  -3373   -604  -3255  -2515    -62  -2963   2976    358  -3207  -3275  -26
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
      5    853   -549  -1759  -1409  -1479  -1119  -1149   -679  -1248  -1287   -615  -1064  -1643  -10
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
      6    361   -484  -2726  -2140   1145  -2255  -1154    698  -1825   1276    595  -1805  -2259  -14
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
      7  -1533  -1224  -3634  -3059   -270  -3280  -2150    621  -2752   1764   3496  -2843  -3028  -22
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
      8   -621   -843  -1909  -1623   -676  -1738  -1165   -166  -1314   2313     76  -1409  -2072  -12
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
      9  -1826  -1443  -4016  -3443   1431  -3619  -2293   1423  -3150   2427    930  -3200  -3234  -24
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
     10    656   -702  -1364   -981  -1067  -1369   -818   -445   -788    787   -207   -900   2412   -7
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
     11   -727   -669  -2450  -1919   -527  -2208  -1281    646  -1631   1628    425  -1721  -2293  -14
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
     12   2322   -683  -1747  -1778  -2648   2229  -1729  -2316  -1866  -2610  -1790  -1197  -1632  -16
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
     13   -778   -898  -1823  -1222   1229  -2071   -620    -71   -642   1224    360  -1191  -2090   -7
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
     14   -970  -1165  -1975  -1692   -832  -2009  -1344   -312  -1384   1910     -5  -1592   2385  -13
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
     15   -362   -856    554   -443  -1171  -1450   -550   -187   -462   -896   -261   -573  -1690   -3
      -   -149   -500    233     43   -381    399    106   -626    210   -466   -720    275    394
      -    -21  -6672  -7714   -894  -1115   -701  -1378      *      *
```

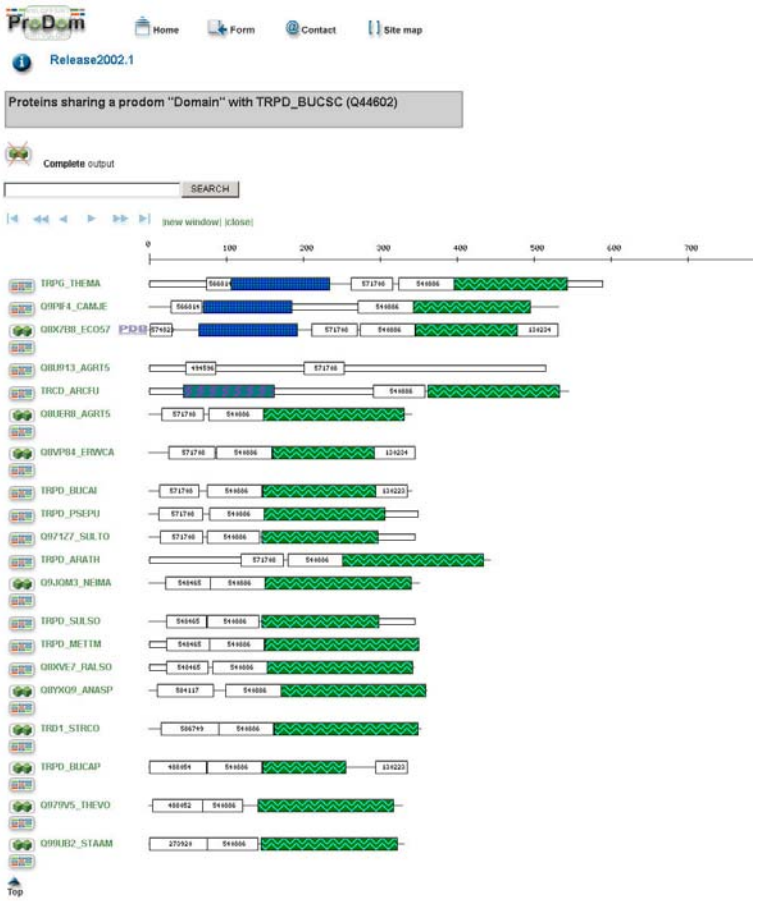
...

...

ProDom

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

- ProDom is a comprehensive set of protein domain families automatically generated from the SWISS-PROT and TrEMBL sequence databases
- consists of an automated compilation of homologous domain alignment.
- 2004.1: ProDom families were generated automatically using PSI-BLAST. built from non fragmentary sequences from SWISS-PROT + TREMBL - Sept, 2003

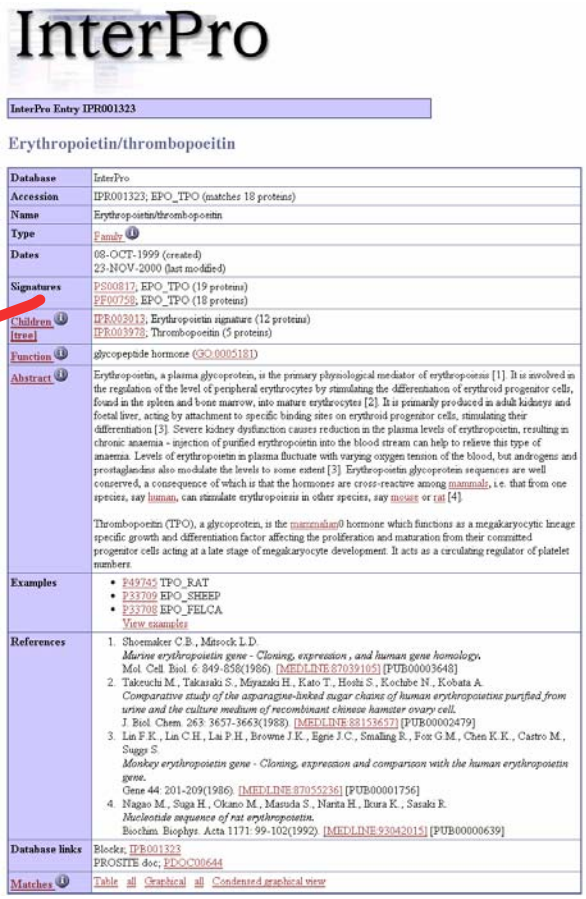


InterPro

www.ebi.ac.uk/interpro

- Search simultaneously many domain databases.
- Single set of documents linked to the various methods;

InterPro release 8.1 contains 11330 entries representing 2933 domains, 8126 families, 222 repeats, 27 active sites, 21 binding sites and 20 post-translational modification sites.



InterPro

InterPro Entry IPR001323

Erythropoietin/thrombopoeitin

Database	InterPro
Accession	IPR001323; EPO_TPO (matches 18 proteins)
Name	Erythropoietin/thrombopoeitin
Type	Family
Dates	08-OCT-1999 (created) 23-NOV-2000 (last modified)
Signatures	PS00817, EPO_TPO (19 proteins) PF00758, EPO_TPO (18 proteins)
Children [tree]	IPR003013, Erythropoietin signature (12 proteins) IPR003978, Thrombopoeitin (5 proteins)
Function	glycopeptide hormone (GO:0005181)
Abstract	Erythropoietin, a plasma glycoprotein, is the primary physiological mediator of erythropoiesis [1]. It is involved in the regulation of the level of peripheral erythrocytes by stimulating the differentiation of erythroid progenitor cells, found in the spleen and bone marrow, into mature erythrocytes [2]. It is primarily produced in adult kidneys and foetal liver, acting by attachment to specific binding sites on erythroid progenitor cells, stimulating their differentiation [3]. Severe kidney dysfunction causes reduction in the plasma levels of erythropoietin, resulting in chronic anaemia - injection of purified erythropoietin into the blood stream can help to relieve this type of anaemia. Levels of erythropoietin in plasma fluctuate with varying oxygen tension of the blood, but androgens and prostaglandins also modulate the levels to some extent [3]. Erythropoietin glycoprotein sequences are well conserved, a consequence of which is that the hormones are cross-reactive among mammals, i.e. that from one species, say human, can stimulate erythropoiesis in other species, say mouse or rat [4]. Thrombopoeitin (TPO), a glycoprotein, is the mammalian0 hormone which functions as a megakaryocytic lineage specific growth and differentiation factor affecting the proliferation and maturation from their committed progenitor cells acting at a late stage of megakaryocyte development. It acts as a circulating regulator of platelet numbers.
Examples	• P49745 TPO_RAT • P33709 EPO_SHEEP • P33708 EPO_FELCA View examples
References	1. Shoemaker C.B., Mitsock L.D. *Murine erythropoietin gene - Cloning, expression, and human gene homology.* Mol. Cell. Biol. 6: 849-858(1986). [MEDLINE:87039105] [PUB00003648] 2. Takeuchi M., Takasaki S., Miyazaki H., Kato T., Hoshi S., Kochibe N., Kobata A. *Comparative study of the asparagine-linked sugar chains of human erythropoietins purified from urine and the culture medium of recombinant chinese hamster ovary cell.* J. Biol. Chem. 263: 3657-3663(1988). [MEDLINE:88158657] [PUB00002479] 3. Lin F.K., Lin C.H., Lai P.H., Browne J.K., Egrie J.C., Smalling R., Fox G.M., Chen K.K., Castro M., Suggs S. *Monkey erythropoietin gene - Cloning, expression and comparison with the human erythropoetin gene.* Gene 44: 201-209(1986). [MEDLINE:87055236] [PUB00001756] 4. Nagao M., Suga H., Okano M., Masuda S., Narita H., Ikura K., Sasaki R. *Nucleotide sequence of rat erythropoietin.* Biochim. Biophys. Acta 1171: 99-102(1992). [MEDLINE:93042015] [PUB00000639]
Database links	Blocks: IPR001323 PROSITE doc; PDOC00644
Matches	Table all Graphical all Condensed graphical view

From a Swiss-Prot entry:

InterPro	IPR007219; Fungal_trans. IPR001138; Fungi_Trscrp_N. IPR005600; Gal4_dimer. Graphical view of domain structure.
Pfam	PF04082; Fungal_trans; 1. PF03902; Gal4_dimer; 1. PF00172; Zn_clus; 1. Pfam graphical view of domain structure.
PRINTS	PR00054; FUNGALZNCYS.
SMART	SM00066; GAL4; 1.
PROSITE	PS00463; ZN2_CY6_FUNGAL_1; 1. PS50048; ZN2_CY6_FUNGAL_2; 1. PROSITE graphical view of domain structure.
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
BLOCKS	P04386.

Example: GAL4_YEAST

InterPro domain architecture:

InterPro Entry	Method accession	Graphical match [?]
IPR001138:	PF00172	
IPR001138:	PR00054	
IPR001138:	PS00463	
IPR001138:	PS50048	
IPR001138:	SM00066	
IPR005600:	PF03902	
IPR007219:	PF04082	
Classification	**PDB Chain/Domain ID & View 3D**	**PDB Chain/Structural Domains [?]**
1d66	1d66a	
1d66	1d66b	
4.10.240.10.1	1d66A0	
g.38.1.1	d1d66a1	
h.1.3.1	d1d66a2	

Categories of databases for Life Sciences

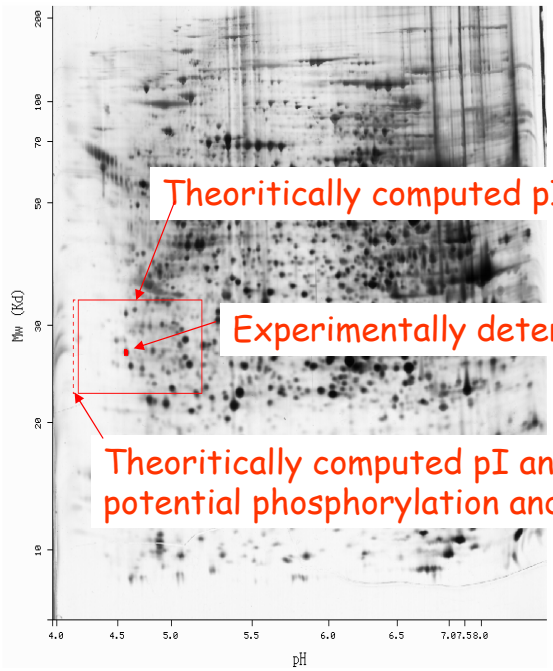
- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

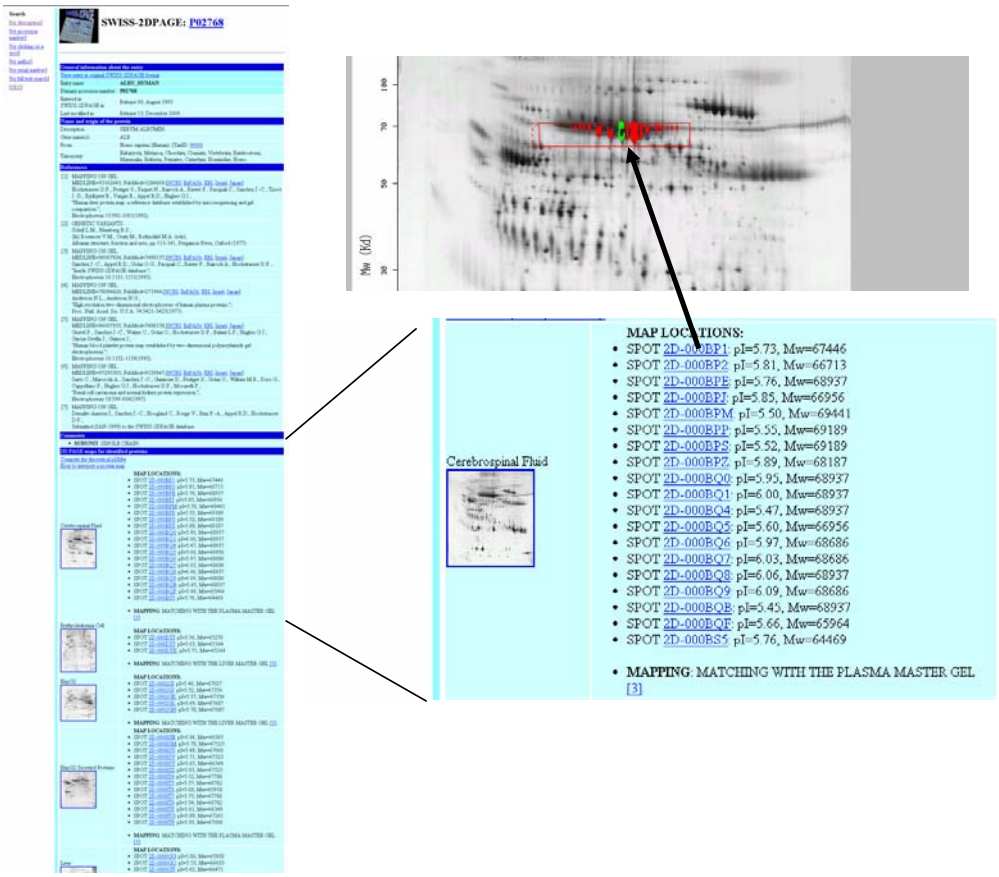
Databases 5: proteomics

- Contain informations obtained by 2D-PAGE: images of master gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Composed of image and text files
- There is currently no protein Mass Spectrometry (MS) database (not for long…)

This protein does not exist in the current release of SWISS-2DPAGE.

SWISS-2DPAGE	GET REGION ON 2D PAGE.





Categories of databases for Life Sciences

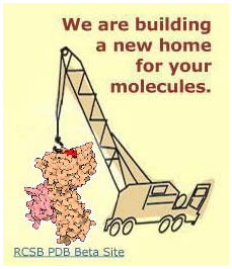
- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Databases 6: 3D structure

- Contain the spatial coordinates of macromolecules whose 3D structure has been obtained by X-ray or NMR studies
- Proteins represent more than 90% of available structures (others are DNA, RNA, sugars, viruses, protein/DNA complexes…)
- Only one: PDB (Protein Data Bank),

PDB: Protein Data Bank

www.rcsb.org/pdb/



- Managed by Research Collaboratory for Structural Bioinformatics (RCSB) (USA).
- Contains structure data on proteins, nucleic acids, protein-nucleic acid complexes, and viruses
- Associated with specialized programs allow the visualization of the corresponding 3D structure (e.g., SwissPDB viewer, Chime, Rasmol)).
- Currently there are ~29'500 structural data for about 8'000 different proteins, but far less protein family (highly redundant) !

PDB: example

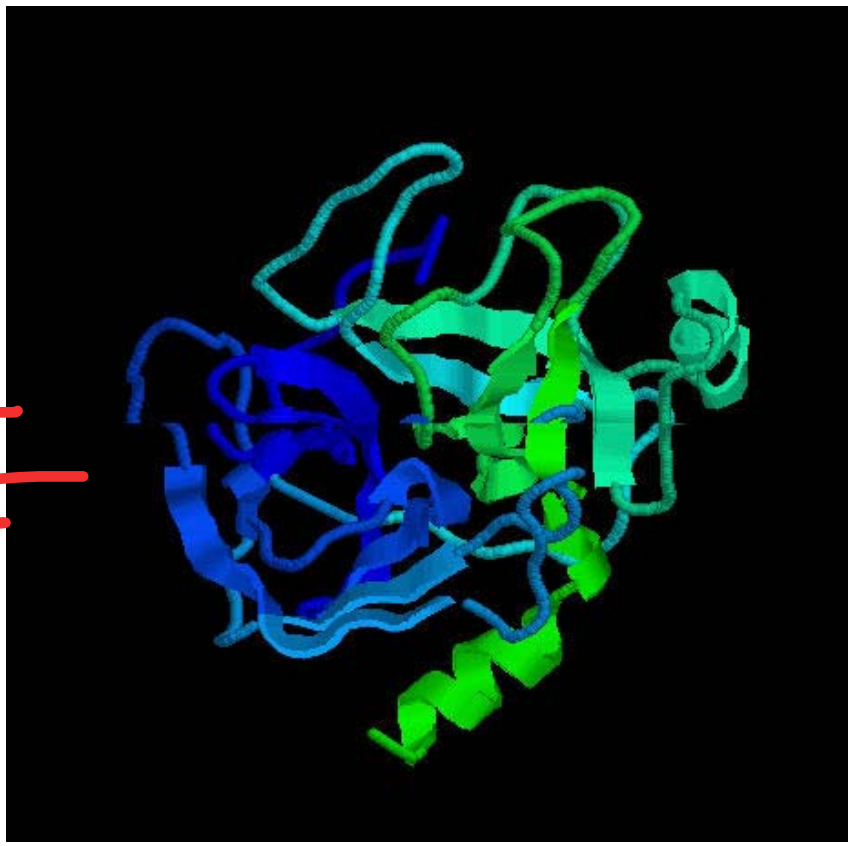
```
HEADER   LYASE(OXO-ACID)                  01-OCT-91  12CA     12CA  2
COMPND   CARBONIC ANHYDRASE /II (CARBONATE DEHYDRATASE) (/HCA II)   12CA  3
COMPND   2 (E.C.4.2.1.1) MUTANT WITH VAL 121 REPLACED BY ALA (/V121A)  12CA  4
SOURCE   HUMAN (HOMO SAPIENS) RECOMBINANT PROTEIN                  12CA  5
AUTHOR   S.K.NAIR,D.W.CHRISTIANSON                                 12CA  6
REVDAT  1  15-OCT-92 12CA   0                                      12CA  7
JRNL     AUTH   S.K.NAIR,T.L.CALDERONE,D.W.CHRISTIANSON,C.A.FIERKE  12CA  8
JRNL     TITL   ALTERING THE MOUTH OF A HYDROPHOBIC POCKET.        12CA  9
JRNL     TITL 2 STRUCTURE AND KINETICS OF HUMAN CARBONIC ANHYDRASE  12CA  10
JRNL     TITL 3 /II$ MUTANTS AT RESIDUE VAL-121                    12CA  11
JRNL     REF    J.BIOL.CHEM.              V. 266 17320 1991     12CA  12
JRNL     REFN   ASTM JBCHA3  US ISSN 0021-9258              071 12CA  13
REMARK  1                                                      12CA  14
REMARK  2                                                      12CA  15
REMARK  2 RESOLUTION. 2.4  ANGSTROMS.                          12CA  16
REMARK  3                                                      12CA  17
REMARK  3 REFINEMENT.                                          12CA  18
REMARK  3   PROGRAM              PROLSQ                        12CA  19
REMARK  3   AUTHORS              HENDRICKSON,KONNERT           12CA  20
REMARK  3   R VALUE             0.170                         12CA  21
REMARK  3   RMSD BOND DISTANCES      0.011  ANGSTROMS          12CA  22
REMARK  3   RMSD BOND ANGLES        1.3    DEGREES            12CA  23
REMARK  4                                                      12CA  24
REMARK  4 N-TERMINAL RESIDUES SER 2, HIS 3, HIS 4 AND C-TERMINAL      12CA  25
REMARK  4 RESIDUE LYS 260 WERE NOT LOCATED IN THE DENSITY MAPS AND,   12CA  26
REMARK  4 THEREFORE, NO COORDINATES ARE INCLUDED FOR THESE RESIDUES.  12CA  27
.....
```

PDB (cont.)

```
SHEET    3   S10 PHE    66  PHE    70 -1  O  ASN    67   N  LEU    60   12CA  68
SHEET    4   S10 TYR    88  TRP    97 -1  O  PHE    93   N  VAL    68   12CA  69
SHEET    5   S10 ALA   116  ASN   124 -1  O  HIS   119   N  HIS    94   12CA  70
SHEET    6   S10 LEU   141  VAL   150 -1  O  LEU   144   N  LEU   120   12CA  71
SHEET    7   S10 VAL   207  LEU   212  1  O  ILE   210   N  GLY   145   12CA  72
SHEET    8   S10 TYR   191  GLY   196 -1  O  TRP   192   N  VAL   211   12CA  73
SHEET    9   S10 LYS   257  ALA   258 -1  O  LYS   257   N  THR   193   12CA  74
SHEET   10   S10 LYS    39  TYR    40  1  O  LYS    39   N  ALA   258   12CA  75
TURN     1  T1 GLN    28  VAL    31     TYPE VIB (CIS-PRO 30)          12CA  76
TURN     2  T2 GLY    81  LEU    84     TYPE II(PRIME) (GLY 82)        12CA  77
TURN     3  T3 ALA   134  GLN   137     TYPE I (GLN 136)               12CA  78
TURN     4  T4 GLN   137  GLY   140     TYPE I (ASP 139)               12CA  79
TURN     5  T5 THR   200  LEU   203     TYPE VIA (CIS-PRO 202)         12CA  80
TURN     6  T6 GLY   233  GLU   236     TYPE II (GLY 235)              12CA  81
CRYST1  42.700   41.700   73.000  90.00 104.60  90.00 P 21         2  12CA  82
ORIGX1      1.000000  0.000000  0.000000        0.00000               12CA  83
ORIGX2      0.000000  1.000000  0.000000        0.00000               12CA  84
ORIGX3      0.000000  0.000000  1.000000        0.00000               12CA  85
SCALE1      0.023419  0.000000  0.006100        0.00000               12CA  86
SCALE2      0.000000  0.023981  0.000000        0.00000               12CA  87
SCALE3      0.000000  0.000000  0.014156        0.00000               12CA  88
ATOM     1  N   TRP     5       8.519  -0.751  10.738  1.00 13.37     12CA  89
ATOM     2  CA  TRP     5       7.743  -1.668  11.585  1.00 13.42     12CA  90
ATOM     3  C   TRP     5       6.786  -2.502  10.667  1.00 13.47     12CA  91
ATOM     4  O   TRP     5       6.422  -2.085   9.607  1.00 13.57     12CA  92
ATOM     5  CB  TRP     5       6.997  -0.917  12.645  1.00 13.34     12CA  93
ATOM     6  CG  TRP     5       5.784  -0.209  12.221  1.00 13.40     12CA  94
ATOM     7  CD1 TRP     5       5.681   1.084  11.797  1.00 13.29     12CA  95
ATOM     8  CD2 TRP     5       4.417  -0.667  12.221  1.00 13.34     12CA  96
ATOM     9  NE1 TRP     5       4.388   1.418  11.515  1.00 13.30     12CA  97
ATOM    10  CE2 TRP     5       3.588   0.375  11.797  1.00 13.35     12CA  98
ATOM    11  CE3 TRP     5       3.837  -1.877  12.645  1.00 13.39     12CA  99
ATOM    12  CZ2 TRP     5       2.216   0.208  11.656  1.00 13.39     12CA 100
ATOM    13  CZ3 TRP     5       2.465  -2.043  12.504  1.00 13.33     12CA 101
ATOM    14  CH2 TRP     5       1.654  -1.001  12.009  1.00 13.34     12CA 102
.....
```

Coordinates of each atom

The same PDB entry "visualized" with Chime

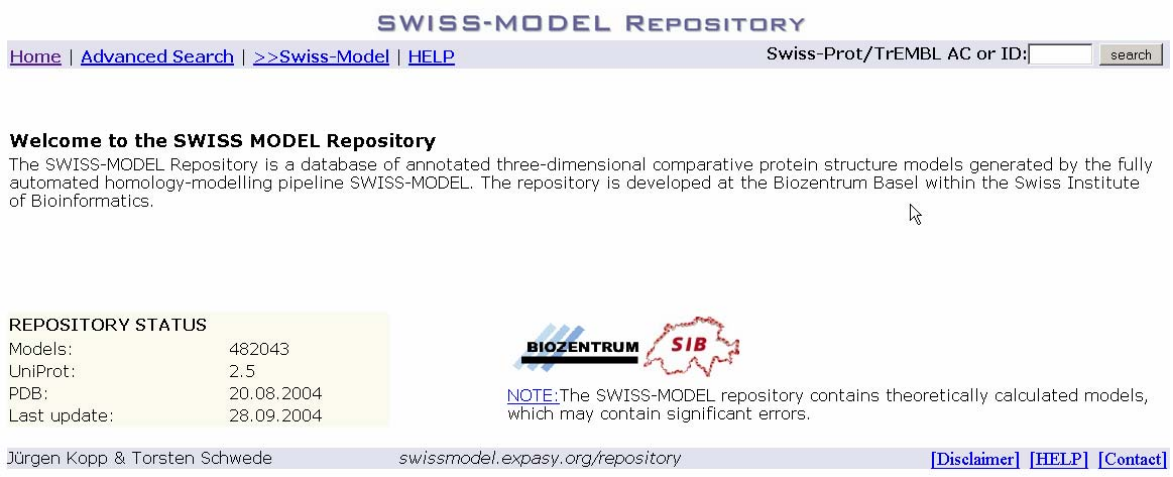


Industry of databases around PDB

- HSSP: Homology-derived secondary structure of proteins.
http://www.sander.ebi.ac.uk/hssp/

- Structure classification
  - -CATH
  - -SCOP
  - -...

- Homology-derived 3D structure db:
Swiss-Model Redepository (SMR): feb 2005: 555'900 models.



Annotated 3D comparative protein structure models generated by the fully automated homology-modelling pipeline SWISS-MODEL.

Precompute the 3D model of protein domains (~200 amino acids, biggest model: 1500 aa) which share about 40 % similarity with a 3D experimentally determined template.

http://swissmodel.expasy.org/repository/

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

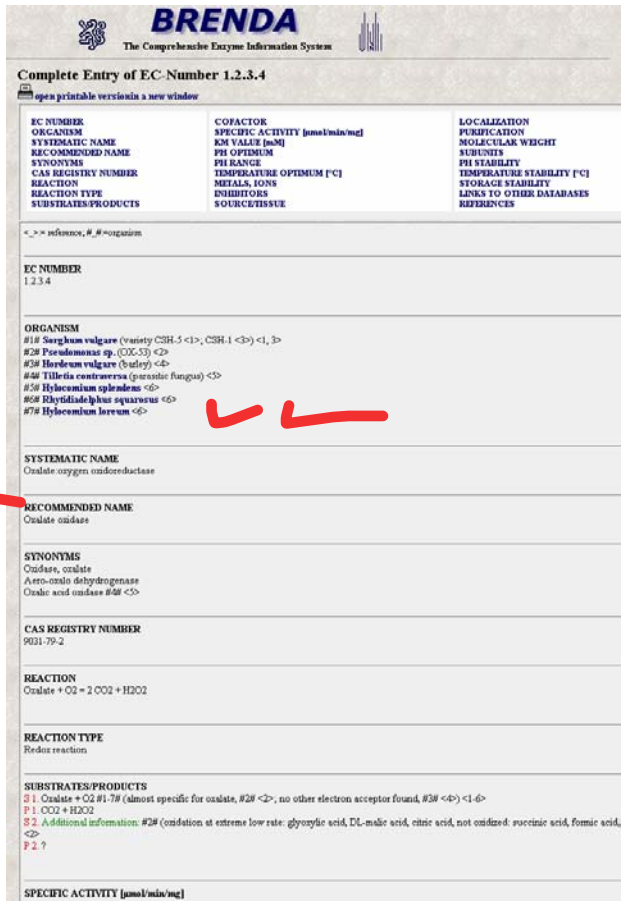
Databases 7: metabolic

- Contain informations that describe enzymes, biochemical reactions and metabolic pathways;
- ENZYME and BRENDA: nomenclature databases that store informations on enzyme names and reactions;
- Metabolic databases: EcoCyc (specialized on Escherichia coli), KEGG, EMP/WIT;
  Usually these databases are tightly coupled with query software that allows the user to visualise reaction schemes.

- There are about 3750 "EC numbers"
  - ~ 1900 are linked to Swiss-Prot sequence
  - ~ 200 are linked to a TrEMBL sequence
  - ~ 1450 can not be linked to any sequence !

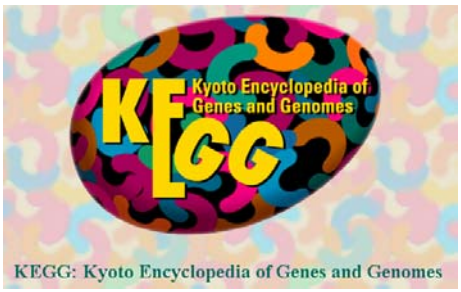
BRENDA
Useful to prepare lab's experiments !

http://www.brenda.uni-koeln.de/

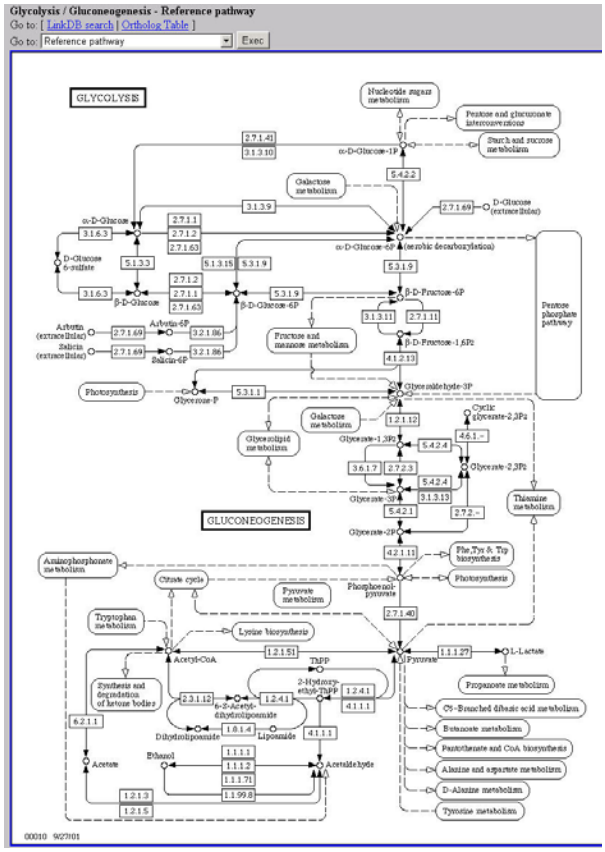


IntEnz = Enzyme + BRENDA + NC-IUBMB nomenclature

http://www.ebi.ac.uk/intenz/index.html



http://www.genome.ad.jp/kegg



Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Databases 8: bibliographic

- Bibliographic reference databases contain citations and abstract informations of published life science articles;
- Example: Medline
- Other more specialized databases also exist (i.e. Agricola http://agricola.nal.usda.gov/, EMBASE (not free)…).

Medline

- Comprehensive database of primary scientific literature in the biomedical area.
- More than 4,000 biomedical journals published in the United States and 70 other countries
- Contains over 15 million indexed citations since 1966 until now
- Citations prior to the mid-1960s are located in OLDMEDLINE.
- Contains links to biological db
  - Many papers not dealing with humans are not in Medline !
  - Before 1970, keeps only the first 10 authors !
  - Not all journals have citations since 1966 ! (they go back…)
  - Indexed by Google in 2004 !

PubMed

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

- Maintained by the US National Library of Medicine.
- Allows access to the citations from MEDLINE and additional life science journals.
- Includes links to many sites providing full text articles and other related resources.
- Gives also access to :
  - In Process Citations
  - Publisher supplied citations: citations directly submitted to PubMed ([Record as supplied by publisher]).
- PMID (PubMed ID) UI (Medline ID)

New:

DOI (Digital Object Identifier) are names (characters and/or digits) assigned to objects of intellectual property such as electronic journal articles, images, learning objects, ebooks, any kind of content.

Server: http://dx.doi.org

-> biggest advance to track documents on the web !

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family (----> tools)
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolism
- Bibliography
- 'Others' (Microarrays, Protein protein interaction…)

Databases 9: others

- There are many databases that cannot be classified in the categories listed previously;
- Examples: ReBase (restriction enzymes), TRANSFAC (transcription factors), CarbBank, GlycoSuiteDB (linked sugars), Protein-protein interactions db (Intact, BIND), Protease db (MEROPS), biotechnology patents db, etc.;
- As well as many other resources concerning any and new aspects of macromolecules and molecular biology (Microarrays).

Amos links: Microarrays

Microarray databases and resources

- ArrayExpress - MicroArray informatics at the EBI
- ChipDB - Gene expression db
- ExpressDB - Yeast and E.coli RNA expression db
- GeneX - Gene Expression project
- GEO - Gene Expression Omnibus
- RAD - RNA abundance db
- SMD - Stanford microarray db

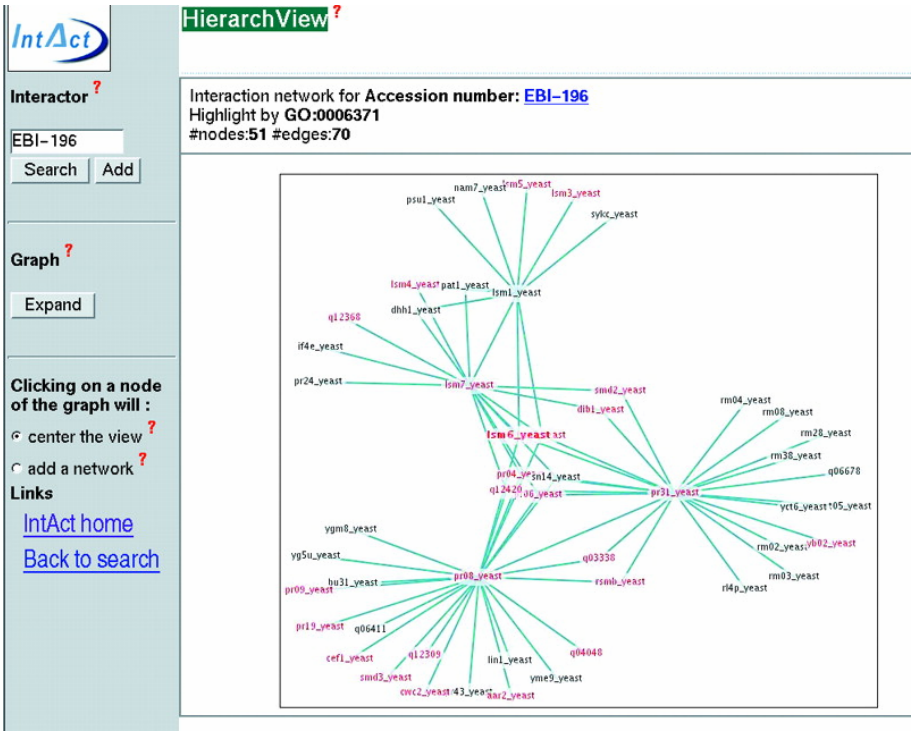
Interactome

- Protein/protein interaction:
description from 1 to more than 20'000 interactions / publication

- Several databases: Intact, BIND, DIP.

- Proteomics standard initiative since 2005

http://www.ebi.ac.uk/intact/index.html



Gene Ontology (GO) database

The Gene Ontology (GO) project (http://www.geneontology.org/) provides structured, controlled vocabularies and classifications that cover several domains of molecular and cellular biology and are freely available for community use in the annotation of genes, gene products and sequences.

The three organizing principles of GO are **molecular function (MF)**, **biological process** (BP) and **cellular component (CC)**.

Proliferation of databases

- Which does contain the highest quality data ?
- Which is the more comprehensive ?
- Which is the more up-to-date ?
- Which is the less redundant ?
- Which is the more indexed (allows complex queries) ?
- Which Web server does respond most quickly ?
- .....?????

To benefit from the data stored in a database, we need:

- easy access to the information

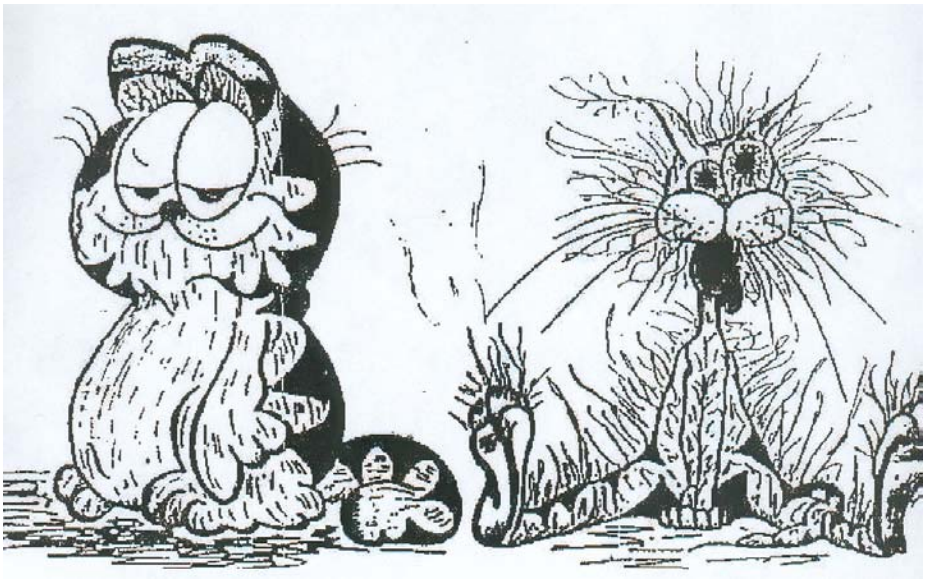
  -> a method for extracting only that information needed to answer a specific biological question

Examples: Entrez (NCBI), SRS (Europ),
tools such as BLAST, Peptident…

Some important practical remarks

- Databases: many errors (automated annotation) !
- Not all db are available on all servers
- The update frequency is not the same for all servers;
- Some servers add automatically cross-references to an entry (implicit links) in addition to already existing links (explicit links)…different looks…

Before the introduction to databases…



After the introduction to databases…