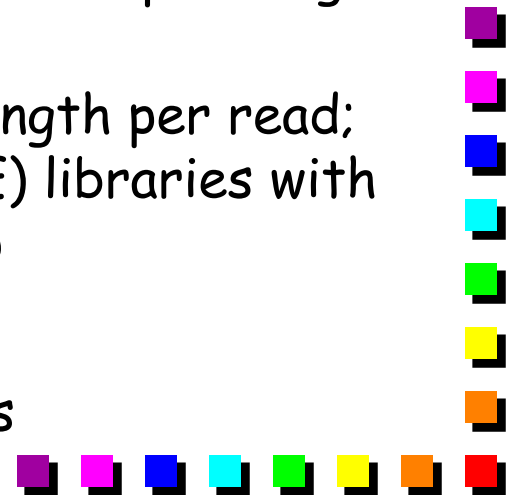


Collecting and storing sequences

- » Collecting sequences
 - »» Storing sequences
 - »»» Using sequences

» Collecting sequences

- How to collect sequences?
 - DNA sequencing by the Sanger method
 - ABI3730: 700-900bp per read
 - DNA sequencing by the high-throughput approaches, the sequencing-by-synthesis (SBS) (pyrosequencing; 454); Solexia
 - 454: 400-500Mb per run and average 450bp in length per read
 - Solexia: ~20Gb per run and 35bp in length per read; paired-end sequencing: paired-end (PE) libraries with inserts of PE ranging from 137-307bp
 - DNA sequencing by chip
 - Genotyping using Illumina technologies

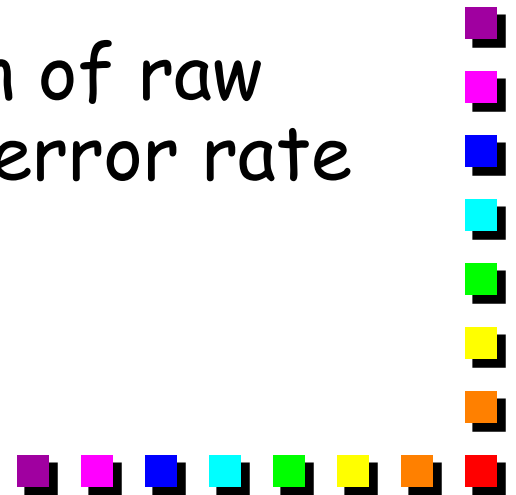


-
- cDNA sequencing of expressed genes by the Sanger and the high-throughput approaches
 - Coding and non-coding genes
 - RNA sequencing by the high-throughput approaches
 - Non-coding small RNAs
 - Protein sequencing by MS etc.



Sequence accuracy varies

- The higher the level of accuracy required in DNA sequences, the more time-consuming and expensive the procedure
- The first-pass sequences
 - EST
 - GSS
 - HTG
- GenBank does permit submission of raw sequence data that have a high error rate
 - An error rate as high as 1 in 100



Submitting sequences to the databases is easy

- The International Nucleotide Sequence Database Collaboration
 - GenBank
 - EMBL
 - DDBJ
- A database accession number is required to publish the sequence
- BankIt: an easy-to-use program with graphic interface for sequence submissions



»»» Storing sequences...

- Introduction
- 2.1 Database : nucleotide sequences
- 2.2 Database : protein sequences
- 2.3 Databases : 'genomics'
- 2.4 Databases : mutation/polymorphism
- 2.5 Database : protein domain/family
- 2.6 Databases : proteomics
- 2.7 Databases : 3D structure
- 2.8 Databases : metabolic
- 2.9 Databases : bibliographic
- 2.10 Databases : others

From Amos Bairoch, et al . SIB, 2002



What is a database ?

- A collection of
 - structured
 - searchable (index) -> table of contents
 - updated periodically (release) -> new edition
 - cross-referenced ([hyperlinks](#)) -> links with other dbdata
- Includes also associated tools (software) necessary for db access, db updating, db information insertion, db information deletion....
- Data storage management: flat files, relational databases...

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.

Some statistics

- 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
- Usually accessible through the web (free !?)
 - Fan's links: <http://ibi.zju.edu.cn/bioinplant>
 - Biohunt: <http://www.expasy.org/BioHunt/>
 - Google: <http://www.google.com/>

■ Some databases in the field of molecular biology...

AATDB, AceDb, ACUTS, ADB, AFDB, AGIS, AMSdb,
ARR, AsDb, BBDB, BCGD, Beanref, Biolmage,
BioMagResBank, BIOMDB, BLOCKS, BovGBASE,
BOVMAP, BSORF, BTKbase, CANSITE, CarbBank,
CARBHYD, CATH, CAZY, CCDC, CD4OLbase, CGAP,
ChickGBASE, Colibri, COPE, CottonDB, CSNDB, CUTG,
CyanoBase, dbCFC, dbEST, dbSTS, DDBJ, DGP, DictyDb,
Picty_cDB, DIP, DOGS, DOMO, DPD, DPlnteract, ECDC,
ECGC, ECO2DBASE, EcoCyc, EcoGene, EMBL, EMD db,
ENZYME, EPD, EpoDB, ESTHER, FlyBase, FlyView,
GCRDB, GDB, GENATLAS, Genbank, GeneCards,
Genline, GenLink, GENOTK, GenProtEC, GIFTS,
GPCRDB, GRAP, GRBase, gRNAsdb, GRR, GSDB,
HAEMB, HAMSTERS, HEART-2DPAGE, HEXAdb, HGMD,
HIDB, HIDC, HIVdb, HotMolecBase, HOVERGEN, HPDB,
HSC-2DPAGE, ICN, ICTVDB, IL2RGbase, IMGT, Kabat,
KDNA, KEGG, Klotho, LGIC, MAD, MaizeDb, MDB,
Medline, Mendel, MEROPS, MGDB, MGI, MHCPEP5
Micado, MitoDat, MITOMAP, MJDB, MmtDB, Mol-R-Us,
MPDB, MRR, MutBase, MycDB, NDB, NRSub, O-lycBase,
OMIA, OMIM, OPD, ORDB, OWL, PAHdb, PatBase, PDB,
PDD, Pfam, PhosphoBase, PigBASE, PIR, PKR, PMD,
PPDB, PRESAGE, PRINTS, ProDom, Prolysis, PROSITE,
PROTOMAP, RatMAP, RDP, REBASE, RGP, SBASE,
SCOP, SeqAnaiRef, SGD, SGP, SheepMap, Soybase,
SPAD, SRNA db, SRPDB, STACK, StyGene, Sub2D,
SubtiList, SWISS-2DPAGE, SWISS-3DIMAGE, SWISS-
MODEL Repository, SWISS-PROT, TelDB, TGN, tmRDB,
TOPS, TRANSFAC, TRR, UniGene, URNADB, V BASE,
VDRR, VectorDB, WDCM, WIT, WormPep, YEPD, YPD,
YPM, etc !!!!

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,...)

Sequences are stored using specified formats

- Ideal minimal content of a « sequence » db
 - Sequences !!
 - Accession number (AC)
 - Taxonomic data
 - References
 - ANNOTATION/CURATION
 - Keywords
 - Cross-references
 - Documentation

Sequence Databases: some « technical » definitions

- Data storage management:
 - flat file: text file
 - relational (e.g., Oracle, Postgres)
 - object oriented (rare in biological field)

- Flat file format:
 - fasta
 - GCG
 - NBRF/PIR
 - MSF....
 - standardized format ?

One sequence format can be converted to another

- READSEQ Online server in [IBI](#), ZJU
- SEQIO for UNIX machines



Examples of entries

- A gene sequence

[AF440696](#) *Arabidopsis thaliana* disease resistance protein RPP4 (RPP4) gene, complete cds

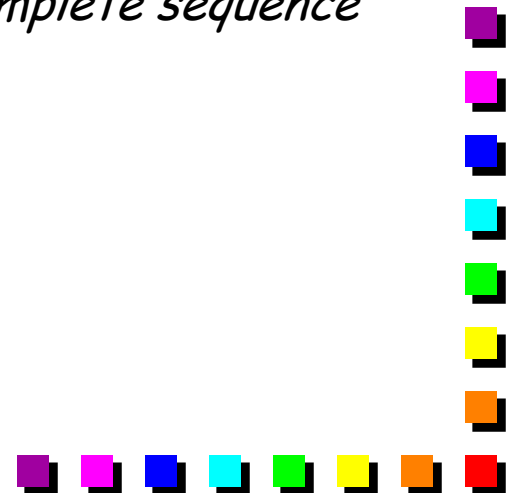
- A genome or chromosome

[NC_002578](#) *Thermoplasma acidophilum*, complete genome;

[NC_003070](#) *Arabidopsis thaliana* chromosome 1, complete sequence

- A protein structure

[1C7K](#) Crystal Structure Of The Zinc Protease



Sequence database : example

SWISS-PROT (flat file)

Accession number	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.
Taxonomy	OS Homo sapiens (Human). OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo. OX NCBI_TaxID=9606; RN [1]
Reference	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 erythropoietin."; RL Nature 313:806-810(1985).
Annotations (comments)	 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).
Cross-references	... 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.
Keywords	 KW Erythrocyte maturation; Glycoprotein; Hormone; Signal; Pharmaceutical.

Sequence database: example (cont.)

Annotations (features)

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 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

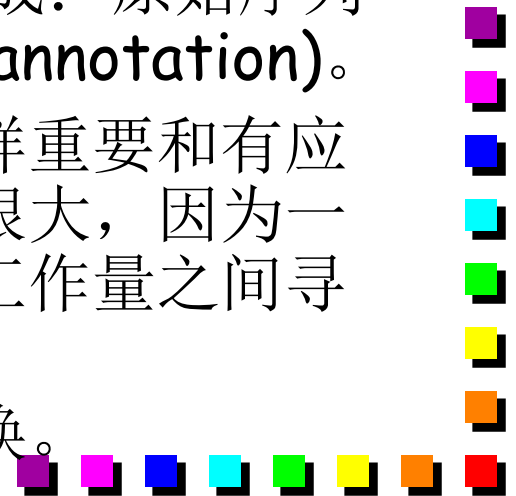
Sequence database: example

...a SWISS-PROT entry, in fasta/Pearson format:

```
>sp|P01588|EPO_HUMAN ERYTHROPOIETIN PRECURSOR - Homo sapiens (Human).  
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE  
NITTGCAEHCSLNENITVPDTKVNIFYAWKRMEVGQQAVEVWQGLALLSEA  
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTLLRALGAQKEAISPPD  
AASAAPLRTITADTFRKLFRVYSNFLRGKCLKLYTGEACRTGDR
```

Summary

- 生物信息学涉及的数据库可大致分为二种：初级数据库(**Primary databases**)和二级数据库(**Secondary databases**)。初级数据库贮存原始的生物数据，如**DNA**序列，由晶体衍射(**Crystallography**)获得的蛋白质结构等。二级数据是在初级数据库的基础上经加工和增加相关信息，如真核生物启动子序列库**EPD**和蛋白质数据库**PROSITE**。
- 一个数据库记录(**entry**)一般由两部分组成：原始序列数据和描述这些数据生物学信息的注释(**annotation**)。
- 注释中包含的信息与相应的序列数据同样重要和有应用价值。不同的数据库的注释质量差异很大，因为一个数据库往往要在数据的完整性和注释工作量之间寻找一个平衡点。
- 各数据库记录的格式有所不同，可以转换。



2.1 Database : nucleotide sequences

- The main DNA sequence db are
EMBL (Europe)/GenBank (USA) /DDBJ (Japan)
- There are also specialized databases for the different types of RNAs (i.e. tRNA, rRNA, tm RNA, uRNA, etc...)
- 3D structure (DNA and RNA)
- Others: Aberrant splicing db; Eucaryotic promoter db (EPD); RNA editing sites, Multimedia Telomere Resource

Nucleotids and associated topics databases (AMOS'links)

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
RNAmdb - 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)

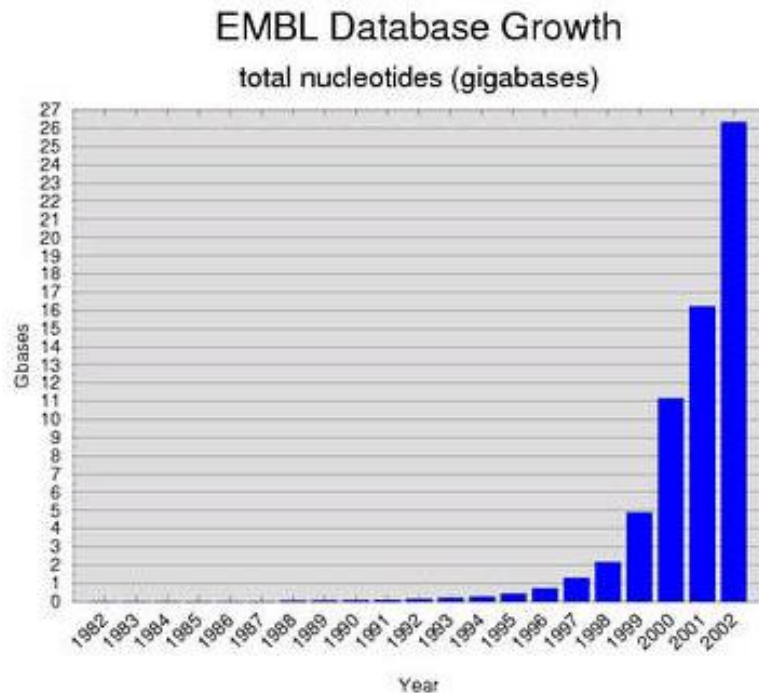
MPDB - Molecular probe db
OPD - Oligonucleotide probe db
VectorDB - Vector sequence db (seems dead!)

EMBL/GenBank/DDBJ

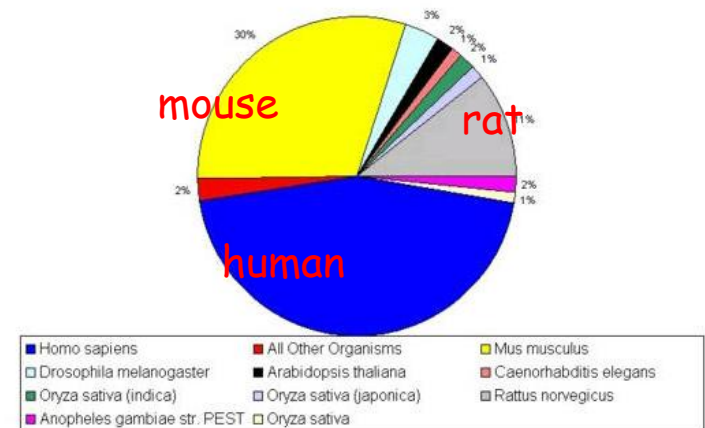
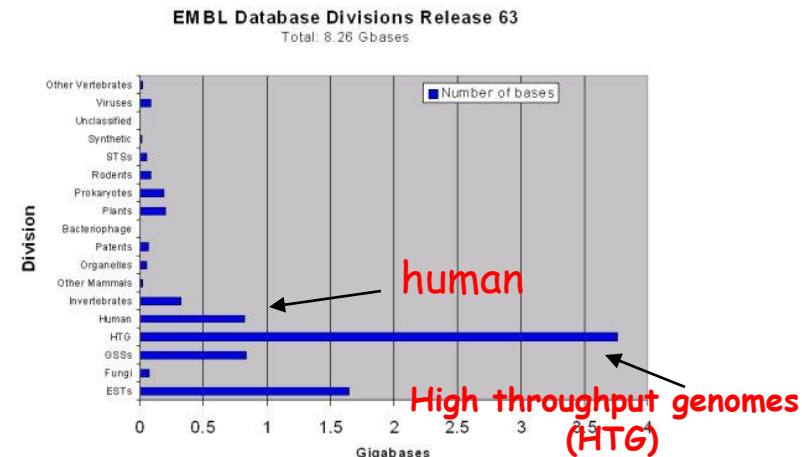
- These 3 db contain mainly the same informations within 2-3 days (few differences in the format and syntax)
- Serve as **archives** containing all sequences (single genes, ESTs, complete genomes, etc.) derived from:
 - Genome projects
 - Sequencing centers
 - Individual scientists
 - Patent offices (i.e. European Patent Office, EPO)
- Non-confidential data are exchanged daily
- Currently: 18×10^6 sequences, over 20×10^9 bp;
- Over the last 12 months the database size has tripled
- Sequences from > 50'000 different species;

The tremendous increase in nucleotide sequences

- EMBL data...first increase in data due to the PCR development...



1980: 80 genes fully sequenced !



Categories/Qualities of nucleotid sequences

ESTs: single pass cDNA reads

GSS: Genome Survey Sequences
single pass genomic DNA sequences

HTG: 'Unfinished' DNA sequences generated by the high-throughput sequencing centers

EMBL/GenBank/DDBJ

- Heterogeneous sequence length: genomes, variants, fragments...
- Sequence sizes:
 - max 300'000 bp /entry (! genomic sequences, overlapping)
 - min 10 bp /entry
- **Archive:** nothing goes out -> **highly redundant !**
- full of errors: in sequences, in annotations, in CDS attribution....
- no consistency of annotations; most annotations are done by the submitters; heterogeneity of the quality and the completion and updating of the informations



- [Index](#)
- [Access](#)
- [Documentation](#)
- [News](#)
- [Submission](#)
- [Group Info](#)
- [Contact](#)

EMBL Submit to the Database

Submission of sequence information to the nucleotide sequence database prior to publication has become standard practice. A unique accession number is assigned by the database which permanently identifies the sequence submitted. The database accession number should be included in the manuscript, preferably on the first page of the journal article, or as required by individual journal procedures. This procedure ensures availability and distribution of new sequence data in a timely fashion.



The preferred submission tool is webin

- [Check sequence for vector contamination \(help\)](#)
- [Webin - WWW Submission System \(help\)](#)
- [Updating EMBL Entries \(help\)](#)
- [Webin-Align \(help\)](#)

Documentation

- [Information for Submitters](#) (including alignments, genome projects and bulk submissions)
- [The DDBJ/EMBL/GenBank Feature Table Definition](#)
- [Sequin](#) (including instructions for downloading)

EMBL entry: example

ID HSERPG standard; 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., Sehra 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

...

keyword

taxonomy

references

Cross-references

EMBL entry (cont.)

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 AKEAENITTGCAEHCSLNENITVPDTKVNIFYAWKRMEVGQQAVEVWQGLALLSEAVLRG

FT QALLVNSSQPWEPLQLHVDKAVSGLRSLTTLRALGAQKEAISPPDAASAAPLRTITAD

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 gcaaccagg catctctgag 60

tctccgccca agaccgggat gccccagg aggtgtccgg gagcccagcc ttccagat 120

CDS
Coding sequence

annotation

sequence

GenBank entry: same entry

LOCUS HSERPG 3398 bp DNA PRI 22-JUN-1993
DEFINITION Human gene for erythropoietin.
ACCESSION X02158
VERSION X02158.1 GI:31224
KEYWORDS erythropoietin; glycoprotein hormone; hormone; signal peptide.
SOURCE human.
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria;
Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 3398)
AUTHORS Jacobs,K., Shoemaker,C., Rudersdorf,R., Neill,S.D., Kaufman,R.J.,
Mufson,A., Seehra,J., Jones,S.S., Hewick,R., Fritsch,E.F.,
Kawakita,M., Shimizu,T. and Miyake,T.
TITLE Isolation and characterization of genomic and cDNA clones of human
erythropoietin
JOURNAL Nature 313 (6005), 806-810 (1985)
MEDLINE 85137899
COMMENT Data kindly reviewed (24-FEB-1986) by K. Jacobs.
FEATURES Location/Qualifiers
source 1..3398
/organism="Homo sapiens"
/db_xref="taxon:9606"
mRNA join(397..627,1194..1339,1596..1682,2294..2473,2608..3327)
exon 397..627
/number=1
sig_peptide join(615..627,1194..1261)
CDS join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
/codon_start=1
/product="erythropoietin"
/protein_id="CAA26095.1"
/db_xref="GI:312304"
/db_xref="SWISS-PROT:P01588"
/translation="MGVHECPAWLWLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL
EAKEAENITTGCAEHCSLNENITVPDTKVNIFYAWKRMEVGQQAVEVWQGLALLSEAVL
RGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLRLALGAQKEAISPPDAASAAPLRTI
..."

GenBank entry (cont.)

```
...
                                TADTFRKLFRVYSNFLRGKCLKLYTGEACRTGDR"
intron      628..1193
            /number=1
exon        1194..1339
            /number=2
mat_peptide join(1262..1339,1596..1682,2294..2473,2608..2760)
            /product="erythropoietin"
intron      1340..1595
            /number=2
exon        1596..1682
            /number=3
intron      1683..2293
            /number=3
exon        2294..2473
            /number=4
intron      2474..2607
            /number=4
exon        2608..3327
            /note="3' untranslated region"
            /number=5
BASE COUNT  698 a  1034 c   991 g   675 t
ORIGIN
    1 agcttctggg cttccagacc cagctacttt gcggaactca gcaacccagg catctctgag
   61 tctccgcccc agaccgggat gccccccagg aggtgtccgg gagcccagcc ttccccagat
  121 agcagctccg ccagtcccaa ggggtgcgca ccggtgcac tcccccccg cgacccaggg
  181 cccgggagca gccccatga cccacacgca cgtctgcagc agccccgtca gccccggagc
  241 ctcaaccagc gcgtcctgcc cctgctctga ccccggtgg cccctacccc tggcgacccc
```

EMBL: The Genome divisions

<http://www.ebi.ac.uk/genomes/>



[EBI Home](#)

[About EBI](#)

[Research](#)

[Services](#)

[Toolbox](#)

[Databases](#)

[Downloads](#)

[Submissions](#)

GENOMES



- [Archaea](#)
- [Bacteria](#)
- [Eukaryota](#)
- [Organelles](#)
- [Phages](#)
- [Plasmids](#)
- [Viroids](#)
- [Viruses](#)

Other Links

- [Proteomes](#)
- [Fasta3 Server](#)
- [Genomes MOT](#)

Completed genomes: EUKARYOTA

Description	Chromosomes	Proteins
Arabidopsis thaliana complete genome (source MIPS)	I II III IV V	I II III IV V
Caenorhabditis elegans complete genome	I-V, X	FastA
Drosophila melanogaster complete genome	X, 2-4, Y	FastA
Encephalitozoon cuniculi complete genome	I II III IV V VI VII VIII IX X XI	I II III IV V VI VII VIII IX X XI
Homo sapiens complete genome: Ensembl project data	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y	
Saccharomyces cerevisiae strain S288C complete genome	I-XVI	FastA

[Schizosaccharomyces pombe strain 972h- complete genome](#)

Human Genome Resources

Genome Project

Sequencing centers

Current progress

Access to data

MapView

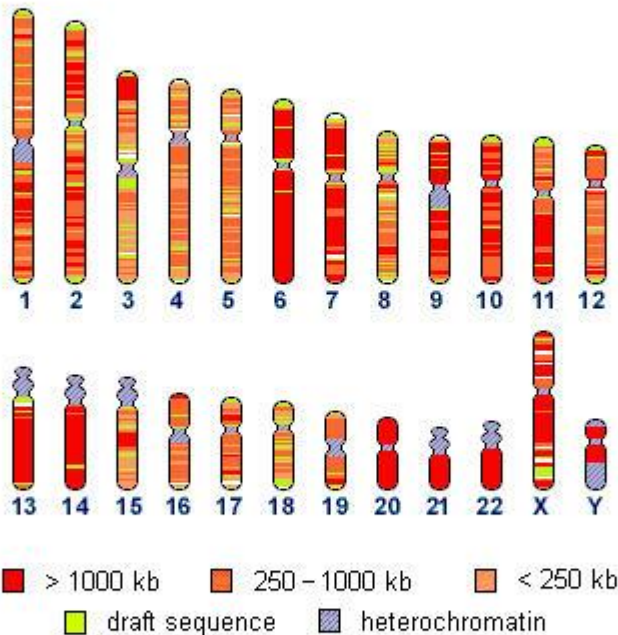
BLAST

Analysis

ePCR

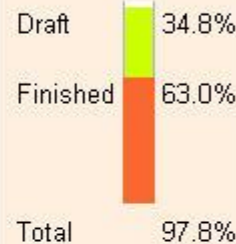
Human Genome Sequencing

Sequencing Progress



Genome Watch

31 Dec, 2001



Assumptions and additional statistics...

What's New

26 June 2000 The International Human Genome Sequencing Consortium Announces the "Working Draft" of the Human Genome.
[More...](#)

Chromosome 21



9 May, 2000

Finished: The clone insert is contiguously sequenced with high quality standard of error rate of 0.01%. There are usually no gaps in the sequence.

Nucleotid databases and « associated » genomic projects/databases

Problem:

Redundancy = makes Blasts searches of the complete databases useless for detecting anything behind the closest homologs.

Solutions:

- assemblies of genomic sequence data (contigs) and corresponding RNA and protein sequences -> dataset of genomic contigs, RNAs and proteins
- annotation of genes, RNAs, proteins, variation (SNPs), STS markers, gene prediction, nomenclature and chromosomal location.
- compute connexions to other resources (cross-references)

Examples: **RefSeq/Locus link** (drosophila, human, mouse, rat and zebrafish), **TIGR** (microbes and plants), **EnsEMBL** (Eukaryota)...

LocusLink / RefSeq
Erythropoietin receptor

2.2 Database : 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
- 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.

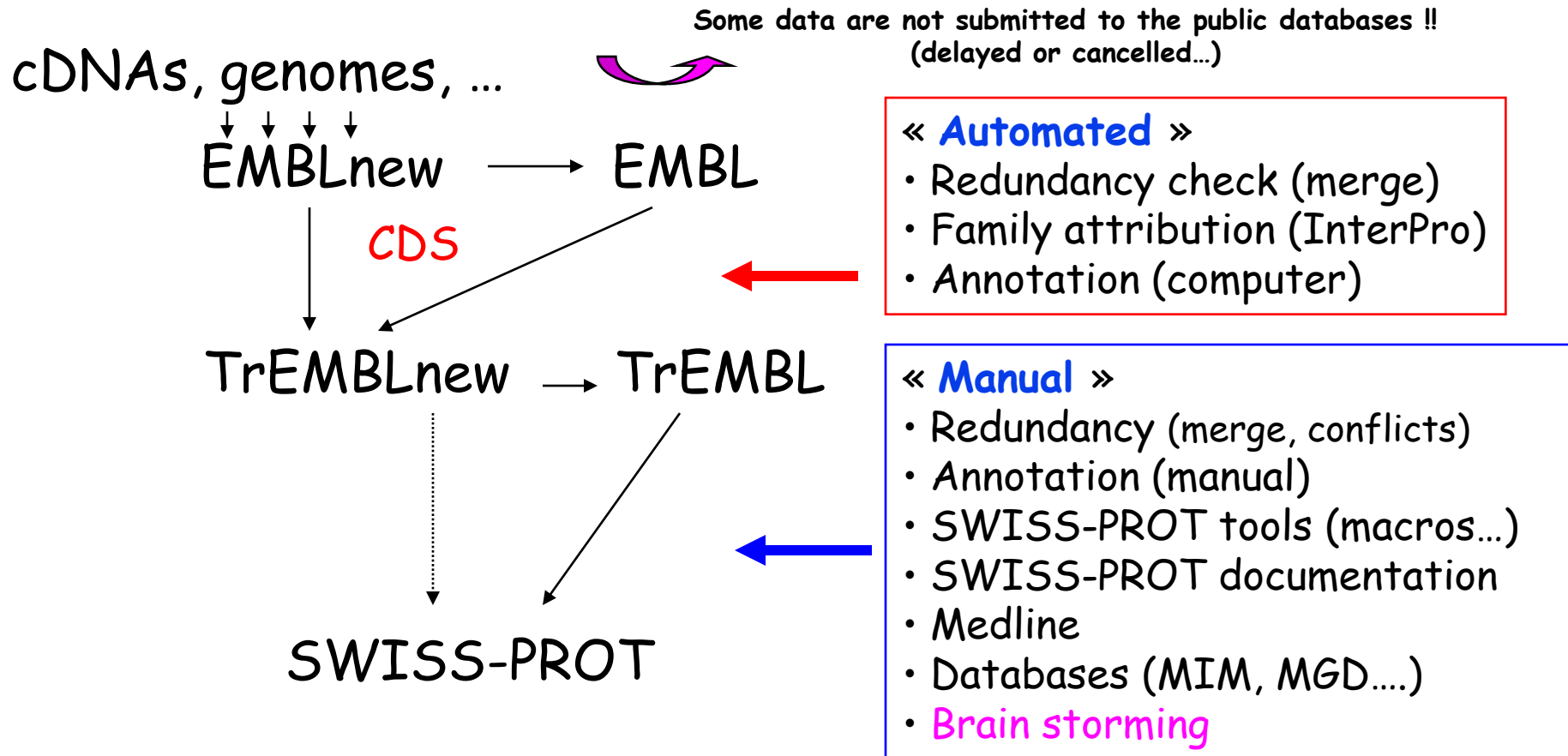
SWISS-PROT

- Collaboration between the SIB (CH) and EMBL/EBI (UK)
- Fully annotated (manually), **non-redundant**, cross-referenced, documented protein sequence database.
- ~113 '000 sequences from more than 6'800 different species; 70 '000 references (publications); 550 '000 cross-references (databases); ~200 Mb of annotations.
- Weekly releases; available from about 50 servers across the world, the main source being ExPASy

TrEMBL (Translation of EMBL)

- It is impossible to cope with the **quantity** of newly generated data AND to maintain the high **quality** of SWISS-PROT -> TrEMBL, created in 1996.
- TrEMBL is automatically generated (from annotated EMBL coding sequences (CDS)) and annotated using software tools.
- Contains all what is **not** in SWISS-PROT.
SWISS-PROT + TrEMBL = all known protein sequences.
- Well-structured SWISS-PROT-like resource.

The simplified story of a SWISS-PROT entry



Once in SWISS-PROT, the entry is no more in TrEMBL, but still in EMBL (archive)

CDS: proposed and submitted at EMBL by authors or by genome projects (can be experimentally proven or derived from gene prediction programs). TrEMBL neither translates DNA sequences, nor uses gene prediction programs: only takes CDS proposed by the submitting authors in the EMBL entry.

	Line code	Content	Occurrence in an entry
	ID	Identification	One; starts the entry
	AC	Accession number(s)	One or more
	DT	Date	Three times
	DE	Description	One or more
	GN	Gene name(s)	Optional
taxonomy	OS	Organism species	One or more
	OG	Organelle	Optional
	OC	Organism classification	One or more
references	RN	Reference number	One or more
	RP	Reference position	One or more
	RC	Reference comment(s)	Optional
	RX	Reference cross-reference(s)	Optional
	RA	Reference authors	One or more
	RT	Reference title	Optional
	RL	Reference location	One or more
	CC	Comments or notes	Optional
	DR	Database cross-references	Optional
	KW	Keywords	Optional
	FT	Feature table data	Optional
	SQ	Sequence header	One
		Amino Acid Sequence	One
	//	Termination line	One; ends the entry

Lines in which you may find 'manual-annotated' information

a Swiss-Prot entry...

overview

SWISS-PROT: O75144

MicePro - a user-friendly view of this SWISS-PROT entry

Entry name
Accession
number

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; Q9HD18;
DT 15-JUL-1999 (Rel. 38, Created)
DT 16-OCT-2001 (Rel. 40, Last sequence update)
DT 01-MAR-2002 (Rel. 41, Last annotation update)
DE ICOS ligand precursor (B7 homolog 2) (B7-H2) (B7-like protein GL50)
DE (B7-related protein-1) (B7RP-1).
GN ICOSL OR B7H2 OR B7RP1 OR KIAA0653.
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. (ISOFORM 1).
RC TISSUE=Dendritic cell;
RX MEDLINE=20477846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
RA Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
RT "Costimulation of T cells by B7-H2, a B7-like molecule that binds
RT ICOS.";
RL Blood 96:2808-2813 (2000).
RN [2]
RP SEQUENCE FROM N.A. (ISOFORM 1), AND CHARACTERIZATION.
RC TISSUE=Peripheral blood lymphocytes;
RX MEDLINE=20465019; PubMed=11007762; [NCBI, ExPASy, EBI, Israel, Japan]
RA Yoshinaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K.,
RA Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H.,
RA Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
RT "Characterization of a new human B7-related protein: B7RP-1 is the
RT ligand to the co-stimulatory protein ICOS.";
RL Int. Immunol. 12:1439-1447 (2000).
RN [3]
RP SEQUENCE FROM N.A. (ISOFORM 2).
RC TISSUE=Leukocyte;
RX MEDLINE=20126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
RA Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A.,
RA Leonard J.P., Hunter S.E., Zoliner R., Thomas J.L., Miyashiro J.S.,
RA Jacobs K.A., Collins M.;
RT "Identification of GL50, a novel B7-like protein that functionally
RT binds to ICOS receptor.";
RL J. Immunol. 164:1653-1657 (2000).
RN [4]
RP SEQUENCE FROM N.A.
RC TISSUE=Brain;
RX MEDLINE=98403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
RA Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,
RA Kotani H., Nomura N., Ohara O.;
RT "Prediction of the coding sequences of unidentified human genes. X.
RT The complete sequences of 100 new cDNA clones from brain which can
RT code for large proteins in vitro.";
RL DNA Res. 5:169-176 (1998).
RN [5]
RP SEQUENCE FROM N.A. (ISOFORM 2).
RA Ling V., Dunussi-Joannopoulos K.;
RT "GL50 molecules and uses thereof.";
RL Patent number: W00121796, 29-MAR-2001.

sequence

CC --!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR
CC ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND
CC CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND
CC DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN
CC MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS
CC WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-
CC STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC --!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC --!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2;
CC ARE PRODUCED BY ALTERNATIVE SPLICING.
CC --!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART,
CC KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE
CC MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES,
CC SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN
CC LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC --!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY
CC TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND
CC MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC --!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG
CC SUBFAMILY.
CC --!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
CC --!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
CC --!- CAUTION: Ref.4 sequence differs from that shown in position 300
CC onward for an unknown reason.
CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration
CC between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC the European Bioinformatics Institute. There are no restrictions on its
CC use by non-profit institutions as long as its content is in no way
CC modified and this statement is not removed. Usage by and for commercial
CC entities requires a license agreement (See <http://www.isb-sib.ch/announce/>
CC or send an email to license@isb-sib.ch).
CC -----
DR EMBL; AF199028; AAF34739.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AB014553; BAA31628.1; ALT SEQ. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR MIM; 605717; -. [NCBI / EBI]
DR GeneCards; ICOSL.
DR GeneLynx; ICOSL.
DR Ensembl; O75144.
DR InterPro; IPR003599; Ig.
DR InterPro; IPR003006; Ig MHC.
DR InterPro; IPR003600; Ig like.
DR InterPro; Graphical view of domain structure.
DR Pfam; PF00047; ig; 3.
DR SMART; SM00409; IG; 1.
DR SMART; SM00410; IG like; 1.
DR ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR BLOCKS; O75144.
DR ProtoMap; O75144.
DR PRESAGE; O75144.
DR DIP; O75144.
DR ModBase; O75144.
DR HUGE; KIAA0653; -.
DR SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW B-cell activation; Immune response; Glycoprotein;
KW Immunoglobulin domain; Signal; Transmembrane; Multigene family;
KW Alternative splicing.
FT SIGNAL 1 18 POTENTIAL.
FT CHAIN 19 302 ICOS LIGAND.
FT DOMAIN 19 256 EXTRACELLULAR (POTENTIAL).
FT TRANSMEM 257 277 POTENTIAL.
FT DOMAIN 278 302 CYTOPLASMIC (POTENTIAL).
FT DOMAIN 30 120 IG-LIKE V-TYPE DOMAIN.
FT DOMAIN 151 223 IG-LIKE C2-TYPE DOMAIN.
FT DISULFID 37 113 POTENTIAL.
FT DISULFID 158 216 POTENTIAL.
FT CARBOHYD 70 70 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 137 137 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 173 173 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 186 186 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 225 225 N-LINKED (GLCNAC...) (POTENTIAL).
FT VARSPIC 300 302 CHL - ESMNLLLS (IN ISOFORM 2)
SQ SEQUENCE 302 AA; 33349 MW; 647934E21B55E34A CRC64;
MRLGSPGLLF LLFSSLRADT QKEVRAWVG SDVELSCACP EGSFRDLNDV YVYVQTSSEK
TVYTHIPQN SSELNVDSRY NRRALMSPAG MLRGDFSRLR FNVTQDEQK FHCILVLSQL
GFQEVLSVEV TLHVAANFSV PVSAPHSPPS QDELTFCTTS INGYPRPNVY WINKNTNSLL
DQALQNDTVF LNMRLGYDVV SVLIARTPS VNIGGCCINLV LQQLNLTGVS QTQNDIGERD
KITENPVSTG ERNAATWSIL AVLCLLVVVA VAIGWVCRDR CLQHSYAGAW AVSPETLTG
RV

SWISS-PROT: O75144

NiceProt - a user-friendly view of this SWISS-PROT entry

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; Q9HD18;
DT 15-JUL-1999 (Rel. 38, Created)
DT 16-OCT-2001 (Rel. 40, Last sequence update)
DT 01-MAR-2002 (Rel. 41, Last annotation update)
DE ICOS ligand precursor (B7 homolog 2) (B7-H2) (B7-like protein GL50)
DE (B7-related protein-1) (B7RP-1).
GN ICOSL_OR_B7H2_OR_B7RP1_OR_KIAA0653.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
CK NCBI TaxID=9606;
RN [1]
RP SEQUENCE FROM N.A. (ISOFORM 1).
RC TISSUE=Dendritic cell;
RX MEDLINE=20477846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
RA Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
RT "Costimulation of T cells by B7-H2, a B7-like molecule that binds
RT ICOS.";
RL Blood 96:2808-2813 (2000).
RN [2]
RP SEQUENCE FROM N.A. (ISOFORM 1), AND CHARACTERIZATION.
RC TISSUE=Peripheral blood lymphocytes;
RX MEDLINE=20465019; PubMed=11007762; [NCBI, ExPASy, EBI, Israel, Japan]
RA Yoshinaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K.,
RA Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H.,
RA Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
RT "Characterization of a new human B7-related protein: B7RP-1 is the
RT ligand to the co-stimulatory protein ICOS.";
RL Int. Immunol. 12:1439-1447 (2000).
RN [3]
RP SEQUENCE FROM N.A. (ISOFORM 2).
RC TISSUE=Leukocyte;
RX MEDLINE=20126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
RA Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A.,
RA Leonard J.P., Hunter S.E., Zoliner R., Thomas J.L., Miyashiro J.S.,
RA Jacobs K.A., Collins M.;
RT "Identification of GL50, a novel B7-like protein that functionally
RT binds to ICOS receptor.";
RL J. Immunol. 164:1653-1657 (2000).
RN [4]
RP SEQUENCE FROM N.A.
RC TISSUE=Brain;
RX MEDLINE=98403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
RA Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,
RA Kotani H., Nomura N., Ohara O.;
RT "Prediction of the coding sequences of unidentified human genes. X.
RT The complete sequences of 100 new cDNA clones from brain which can
RT code for large proteins in vitro.";
RL DNA Res. 5:169-176 (1998).
RN [5]
RP SEQUENCE FROM N.A. (ISOFORM 2).
RA Ling V., Dunussi-Joannopoulos K.;
RT "GL50 molecules and uses thereof.";
RL Patent number W00121796, 29-MAR-2001.

CC -!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR
CC ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND
CC CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND
CC DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN
CC MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS
CC WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-
CC STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC -!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC -!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2;
CC ARE PRODUCED BY ALTERNATIVE SPLICING.
CC -!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART,
CC KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE
CC MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES,
CC SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN
CC LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC -!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY
CC TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND
CC MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC -!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG
CC SUBFAMILY.
CC -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
CC -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
CC -!- CAUTION: Ref.4 sequence differs from that shown in position 300
CC onward for an unknown reason.

CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration
CC between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC the European Bioinformatics Institute. There are no restrictions on its
CC use by non-profit institutions as long as its content is in no way
CC modified and this statement is not removed. Usage by and for commercial
CC entities requires a license agreement (See <http://www.isb-sib.ch/announce/>
CC or send an email to license@isb-sib.ch).
CC -----

DR EMBL; AF199028; AAF34739.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AB014553; BAA31628.1; ALT SEQ. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR MIM; 605717; -. [NCBI / EBI]
DR GeneCards: ICOSL.
DR GeneLynx: ICOSL.
DR Ensembl: O75144.
DR InterPro: IPR003599; Ig.
DR InterPro: IPR003006; Ig MHC.
DR InterPro: IPR003600; Ig like.
DR InterPro: Graphical view of domain structure.
DR Pfam: PF00047; ig; 3.
DR SMART: SM00409; IG; 1.
DR SMART: SM00410; IG like; 1.
DR ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR BLOCKS: O75144.
DR ProtoMap: O75144.
DR PRESAGE: O75144.
DR DIP: O75144.
DR ModBase: O75144.
DR HUGE: KIAA0653; -.
DR SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW B-cell activation; Immune response; Glycoprotein;
KW Immunoglobulin domain; Signal; Transmembrane; Multigene family;
KW Alternative splicing.
FT SIGNAL 1 18 POTENTIAL.
FT CHAIN 19 302 ICOS LIGAND.
FT DOMAIN 19 256 EXTRACELLULAR (POTENTIAL).
FT TRANSMEM 257 277 POTENTIAL.
FT DOMAIN 278 302 CYTOPLASMIC (POTENTIAL).
FT DOMAIN 30 120 IG-LIKE V-TYPE DOMAIN.
FT DOMAIN 151 223 IG-LIKE C2-TYPE DOMAIN.
FT DISULFID 37 113 POTENTIAL.
FT DISULFID 158 216 POTENTIAL.
FT CARBOHYD 70 70 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 137 137 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 173 173 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 186 186 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 225 225 N-LINKED (GLCNAC...) (POTENTIAL).
FT VARSPLIC 300 302 GHV -> ESWNLLLLLS (IN ISOFORM 2).
SQ SEQUENCE 302 AA; 33349 MW; 647934E21B5E34A CRC64;
MRLGSPGLLF LLFSSLRADT QKEVEKRMVG SDVELSCACP EGSFRDLNDV YVWQTSSEK
TVYTHIPQN SSSLNVDSRY NRRALMSPAG MLRGDFSLRL FNVTQDEQK FHCILVLSQL
GFQEVLSVF LNRHGLYDVF PVVSAPHSPPS QDELTTCTTS INGYPRPNVY WINKNTNSLL
DQALQNDTVF LNRHGLYDVF SVLRIARTPS VNIGGTCCTV LLQONLTVGS QTQNDIGERD
KITEPNVSTG EKNAATVSLI EAVCLLVVVA VAIGWVCNDR CLQHSYAGAW AVSPETLGT
HV

SWISS-PROT: O75144

NiceProt - a user-friendly view of this SWISS-PROT entry

```
ID      ICOL_HUMAN      STANDARD;      PRT;      302 AA.
AC      O75144; Q9NRQ1; Q9HD18;
DT      15-JUL-1999 (Rel. 38, Created)
DT      16-OCT-2001 (Rel. 40, Last sequence update)
DT      01-MAR-2002 (Rel. 41, Last annotation update)
DE      ICOS ligand precursor (B7 homolog 2) (B7-H2) (B7-like protein GL50)
DE      (B7-related protein-1) (B7RP-1).
GN      ICOSL OR B7H2 OR B7RP1 OR KIAA0653.
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. (ISOFORM 1).
RC      TISSUE=Dendritic cell;
RX      MEDLINE=20477846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
RT      "Costimulation of T cells by B7-H2, a B7-like molecule that binds
RT      ICOS.";
RL      Blood 96:2808-2813 (2000).
RN      [2]
RP      SEQUENCE FROM N.A. (ISOFORM 1), AND CHARACTERIZATION.
RC      TISSUE=Peripheral blood lymphocytes;
RX      MEDLINE=20465019; PubMed=11007762; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Yoshinaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K.,
RA      Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H.,
RA      Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
RT      "Characterization of a new human B7-related protein: B7RP-1 is the
RT      ligand to the co-stimulatory protein ICOS.";
RL      Int. Immunol. 12:1439-1447 (2000).
RN      [3]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RC      TISSUE=Leukocyte;
RX      MEDLINE=20126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A.,
RA      Leonard J.P., Hunter S.E., Zoliner R., Thomas J.L., Miyashiro J.S.,
RA      Jacobs K.A., Collins M.;
RT      "Identification of GL50, a novel B7-like protein that functionally
RT      binds to ICOS receptor.";
RL      J. Immunol. 164:1653-1657 (2000).
RN      [4]
RP      SEQUENCE FROM N.A.
RC      TISSUE=Brain;
RX      MEDLINE=98403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,
RA      Kotani H., Nomura N., Ohara O.;
RT      "Prediction of the coding sequences of unidentified human genes. X.
RT      The complete sequences of 100 new cDNA clones from brain which can
RT      code for large proteins in vitro.";
RL      DNA Res. 5:169-176 (1998).
RN      [5]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RA      Ling V., Dunussi-Joannopoulos K.;
RT      "GL50 molecules and uses thereof.";
RL      Patent number WOO121796, 29-MAR-2001.
```

References

```
CC      -!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR
CC      ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND
CC      CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND
CC      DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN
CC      MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS
CC      WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-
CC      STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC      -!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC      -!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2;
CC      ARE PRODUCED BY ALTERNATIVE SPLICING.
CC      -!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART,
CC      KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE
CC      MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES,
CC      SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN
CC      LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC      -!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY
CC      TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND
CC      MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC      -!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG
CC      SUBFAMILY.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
CC      -!- CAUTION: Ref.4 sequence differs from that shown in position 300
CC      onward for an unknown reason.
CC      -----
CC      This SWISS-PROT entry is copyright. It is produced through a collaboration
CC      between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC      the European Bioinformatics Institute. There are no restrictions on its
CC      use by non-profit institutions as long as its content is in no way
CC      modified and this statement is not removed. Usage by and for commercial
CC      entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC      or send an email to license@isb-sib.ch).
CC      -----
DR      EMBL; AF199028; AAF34739.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AB014553; BAA31628.1; ALT SEQ. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      MIM; 605717; -. [NCBI / EBI]
DR      GeneCards: ICOSL.
DR      GeneLynx: ICOSL.
DR      Ensembl: O75144.
DR      InterPro: IPR003599; Ig.
DR      InterPro: IPR003006; Ig MHC.
DR      InterPro: IPR003600; Ig like.
DR      InterPro: Graphical view of domain structure.
DR      Pfam: PF00047; ig; 3.
DR      SMART: SM00409; IG; 1.
DR      SMART: SM00410; IG like; 1.
DR      ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR      BLOCKS: O75144.
DR      ProtoMap: O75144.
DR      PRESAGE: O75144.
DR      DIP: O75144.
DR      ModBase: O75144.
DR      HUGE: KIAA0653; -.
DR      SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW      B-cell activation; Immune response; Glycoprotein;
KW      Immunoglobulin domain; Signal; Transmembrane; Multigene family;
KW      Alternative splicing.
FT      SIGNAL      1      18      POTENTIAL.
FT      CHAIN       19      302      ICOS LIGAND.
FT      DOMAIN      19      256      EXTRACELLULAR (POTENTIAL).
FT      TRANSMEM    257      277      POTENTIAL.
FT      DOMAIN      278      302      CYTOPLASMIC (POTENTIAL).
FT      DOMAIN      30      120      IG-LIKE V-TYPE DOMAIN.
FT      DOMAIN      151      223      IG-LIKE C2-TYPE DOMAIN.
FT      DISULFID     37      113      POTENTIAL.
FT      DISULFID    158      216      POTENTIAL.
FT      CARBOHYD     70      70      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    137      137      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    173      173      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    186      186      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    225      225      N-LINKED (GLCNAC...) (POTENTIAL).
FT      VARSPLIC    300      302      GHV -> ESWNLLLLLS (IN ISOFORM 2).
SQ      SEQUENCE 302 AA; 33349 MW; 647934E21B55E34A CRC64;
        MRLGSPGLLF LLFSSLRADT QKEVEKRMVG SDVELSCAP EGSFRDLNDV YVWQTSSEK
        TVVYTHIPQN SSENVDVSRV NRRALMSPAG MLRGDFSRLP FNVTQDEQK FHCILVLSQK
        GFQEVLSVFE LNRHVAIVSFV PVSAPHSFS QDELTTCTTS INGYPRPNVY WINKNTNSLL
        DQALQNDTVF LNRHGLYDVG SVLIARTGPS VNIGGTCVNI LLQQLNLTGVS QTQNDIGERD
        KITEPNVSTG ENNAATVSLI AVLCLLVVVA VAIGWVCNDR CLQHSYAGAW AVSPETELTG
        HV
```

Comments

SWISS-PROT: O75144

NiceProt - a user-friendly view of this SWISS-PROT entry

```
ID      ICOL_HUMAN      STANDARD;      PRT;      302 AA.
AC      O75144; Q9NRQ1; Q9HD18;
DT      15-JUL-1999 (Rel. 38, Created)
DT      16-OCT-2001 (Rel. 40, Last sequence update)
DT      01-MAR-2002 (Rel. 41, Last annotation update)
DE      ICOS ligand precursor (B7 homolog 2) (B7-H2) (B7-like protein GL50)
DE      (B7-related protein-1) (B7RP-1).
GN      ICOSL OR B7H2 OR B7RP1 OR KIAA0653.
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. (ISOFORM 1).
RC      TISSUE=Dendritic cell;
RX      MEDLINE=20477846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
RT      "Costimulation of T cells by B7-H2, a B7-like molecule that binds
RT      ICOS.";
RL      Blood 96:2808-2813 (2000).
RN      [2]
RP      SEQUENCE FROM N.A. (ISOFORM 1), AND CHARACTERIZATION.
RC      TISSUE=Peripheral blood lymphocytes;
RX      MEDLINE=20465019; PubMed=11007762; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Yoshinaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K.,
RA      Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H.,
RA      Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
RT      "Characterization of a new human B7-related protein: B7RP-1 is the
RT      ligand to the co-stimulatory protein ICOS.";
RL      Int. Immunol. 12:1439-1447 (2000).
RN      [3]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RC      TISSUE=Leukocyte;
RX      MEDLINE=20126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A.,
RA      Leonard J.P., Hunter S.E., Zoliner R., Thomas J.L., Miyashiro J.S.,
RA      Jacobs K.A., Collins M.;
RT      "Identification of GL50, a novel B7-like protein that functionally
RT      binds to ICOS receptor.";
RL      J. Immunol. 164:1653-1657 (2000).
RN      [4]
RP      SEQUENCE FROM N.A.
RC      TISSUE=Brain;
RX      MEDLINE=98403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,
RA      Kotani H., Nomura N., Ohara O.;
RT      "Prediction of the coding sequences of unidentified human genes. X.
RT      The complete sequences of 100 new cDNA clones from brain which can
RT      code for large proteins in vitro.";
RL      DNA Res. 5:169-176 (1998).
RN      [5]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RA      Ling V., Dunussi-Joannopoulos K.;
RT      "GL50 molecules and uses thereof.";
RL      Patent number W00121796, 29-MAR-2001.
```

```
CC      -!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR
CC      ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND
CC      CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND
CC      DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN
CC      MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS
CC      WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-
CC      STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC      -!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC      -!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS: 1 (SHOWN HERE) AND 2;
CC      ARE PRODUCED BY ALTERNATIVE SPLICING.
CC      -!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART,
CC      KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE
CC      MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES,
CC      SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN
CC      LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC      -!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY
CC      TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND
CC      MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC      -!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG
CC      SUBFAMILY.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
CC      -!- CAUTION: Ref.4 sequence differs from that shown in position 300
CC      onward for an unknown reason.
```

```
CC      -----
CC      This SWISS-PROT entry is copyright. It is produced through a collaboration
CC      between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC      the European Bioinformatics Institute. There are no restrictions on its
CC      use by non-profit institutions as long as its content is in no way
CC      modified and this statement is not removed. Usage by and for commercial
CC      entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC      or send an email to license@isb-sib.ch).
CC      -----
DR      EMBL; AF199028; AAF34739.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AB014553; BAA31628.1; ALT SEQ. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      MIM; 605717; -. [NCBI / EBI]
DR      GeneCards: ICOSL.
DR      GeneLynx: ICOSL.
DR      Ensembl: O75144.
DR      InterPro: IPR003599; Ig.
DR      InterPro: IPR003006; Ig MHC.
DR      InterPro: IPR003600; Ig like.
DR      InterPro: Graphical view of domain structure.
DR      Pfam: PF00047; ig; 3.
DR      SMART: SM00409; IG; 1.
DR      SMART: SM00410; IG like; 1.
DR      ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR      BLOCKS: O75144.
DR      ProtoMap: O75144.
DR      PRESAGE: O75144.
DR      DIP: O75144.
DR      ModBase: O75144.
DR      HUGE: KIAA0653; -.
DR      SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW      B-cell activation; Immune response; Glycoprotein;
KW      Immunoglobulin domain; Signal; Transmembrane; Multigene family;
KW      Alternative splicing.
FT      SIGNAL      1      18      POTENTIAL.
FT      CHAIN      19      302      ICOS LIGAND.
FT      DOMAIN      19      256      EXTRACELLULAR (POTENTIAL).
FT      TRANSMEM      257      277      POTENTIAL.
FT      DOMAIN      278      302      CYTOPLASMIC (POTENTIAL).
FT      DOMAIN      30      120      IG-LIKE V-TYPE DOMAIN.
FT      DOMAIN      151      223      IG-LIKE C2-TYPE DOMAIN.
FT      DISULFID      37      113      POTENTIAL.
FT      DISULFID      158      216      POTENTIAL.
FT      CARBOHYD      70      70      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      137      137      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      173      173      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      186      186      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      225      225      N-LINKED (GLCNAC...) (POTENTIAL).
FT      VARSPLIC      300      302      GHV -> ESWNLLLLLS (IN ISOFORM 2).
SQ      SEQUENCE      302 AA; 33349 MW; 647934E21B5E34A CRC64;
      MRLGSPGLLF LLFSSLRADT QKEVEKRMVG SDVELSCAP EGSFRDLNDV YVWQTSSEK
      TVYTHIPQN SSSLNVDSRY NRRALMSPAG MLRGDFSRL FNVTQDEQK FHCILVLSQL
      GFQEVLSVEV LNRVAANFSV PVSAPHSPPS QDELTTCTTS INGYPRPNVY WINKNTNSLL
      DQALQNDTVF LNMRLGYDWW SVLIARTGPS VNIGWCINNV LLQONLTVGS QTQNDIGERD
      KITEPNVSTG EKNAATVSLI AVLCLLVVVA VAIGWVCNDR CLQHSYAGAW AVSPETLGT
      HV
```

SWISS-PROT: O75144

NiceProt - a user-friendly view of this SWISS-PROT entry

```
ID      ICOL_HUMAN      STANDARD;      PRT;      302 AA.
AC      O75144; Q9NRQ1; Q9HD18;
DT      15-JUL-1999 (Rel. 38, Created)
DT      16-OCT-2001 (Rel. 40, Last sequence update)
DT      01-MAR-2002 (Rel. 41, Last annotation update)
DE      ICOS ligand precursor (B7 homolog 2) (B7-H2) (B7-like protein GL50)
DE      (B7-related protein-1) (B7RP-1).
GN      ICOSL OR B7H2 OR B7RP1 OR KIAA0653.
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. (ISOFORM 1).
RC      TISSUE=Dendritic cell;
RX      MEDLINE=20477846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
RT      "Costimulation of T cells by B7-H2, a B7-like molecule that binds
RT      ICOS.";
RL      Blood 96:2808-2813 (2000).
RN      [2]
RP      SEQUENCE FROM N.A. (ISOFORM 1), AND CHARACTERIZATION.
RC      TISSUE=Peripheral blood lymphocytes;
RX      MEDLINE=20465019; PubMed=11007762; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Yoshinaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K.,
RA      Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H.,
RA      Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
RT      "Characterization of a new human B7-related protein: B7RP-1 is the
RT      ligand to the co-stimulatory protein ICOS.";
RL      Int. Immunol. 12:1439-1447 (2000).
RN      [3]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RC      TISSUE=Leukocyte;
RX      MEDLINE=20126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A.,
RA      Leonard J.P., Hunter S.E., Zoliner R., Thomas J.L., Miyashiro J.S.,
RA      Jacobs K.A., Collins M.;
RT      "Identification of GL50, a novel B7-like protein that functionally
RT      binds to ICOS receptor.";
RL      J. Immunol. 164:1653-1657 (2000).
RN      [4]
RP      SEQUENCE FROM N.A.
RC      TISSUE=Brain;
RX      MEDLINE=98403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,
RA      Kotani H., Nomura N., Ohara O.;
RT      "Prediction of the coding sequences of unidentified human genes. X.
RT      The complete sequences of 100 new cDNA clones from brain which can
RT      code for large proteins in vitro.";
RL      DNA Res. 5:169-176 (1998).
RN      [5]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RA      Ling V., Dunussi-Joannopoulos K.;
RT      "GL50 molecules and uses thereof.";
RL      Patent number: WOO121796, 29-MAR-2001.
```

Cross-references

```
CC      -!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR
CC      ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND
CC      CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND
CC      DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN
CC      MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS
CC      WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-
CC      STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC      -!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC      -!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2;
CC      ARE PRODUCED BY ALTERNATIVE SPLICING.
CC      -!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART,
CC      KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE
CC      MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES,
CC      SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN
CC      LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC      -!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY
CC      TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND
CC      MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC      -!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG
CC      SUBFAMILY.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
CC      -!- CAUTION: Ref.4 sequence differs from that shown in position 300
CC      onward for an unknown reason.
CC      -----
CC      This SWISS-PROT entry is copyright. It is produced through a collaboration
CC      between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC      the European Bioinformatics Institute. There are no restrictions on its
CC      use by non-profit institutions as long as its content is in no way
CC      modified and this statement is not removed. Usage by and for commercial
CC      entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC      or send an email to license@isb-sib.ch).
CC      -----
DR      EMBL; AF199028; AAF34739.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AB014553; BAA31628.1; ALT. SEQ. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      MIM; 605717; -. [NCBI / EBI]
DR      GeneCards: ICOSL.
DR      GeneLynx: ICOSL.
DR      Ensembl: O75144.
DR      InterPro: IPR003599; Ig.
DR      InterPro: IPR003006; Ig MHC.
DR      InterPro: IPR003600; Ig like.
DR      InterPro: Graphical view of domain structure.
DR      Pfam: PF00047; ig; 3.
DR      SMART: SM00409; IG; 1.
DR      SMART: SM00410; IG like; 1.
DR      ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR      BLOCKS: O75144.
DR      ProtoMap: O75144.
DR      PRESAGE: O75144.
DR      DIP: O75144.
DR      ModBase: O75144.
DR      HUGE: KIAA0653; -.
DR      SWISS-2DPAGE: GET REGION ON 2D PAGE.
KW      B-cell activation; Immune response; Glycoprotein;
KW      Immunoglobulin domain; Signal; Transmembrane; Multigene family;
KW      Alternative splicing.
FT      SIGNAL      1      18      POTENTIAL.
FT      CHAIN      19      302      ICOS LIGAND.
FT      DOMAIN      19      256      EXTRACELLULAR (POTENTIAL).
FT      TRANSMEM      257      277      POTENTIAL.
FT      DOMAIN      278      302      CYTOPLASMIC (POTENTIAL).
FT      DOMAIN      30      120      IG-LIKE V-TYPE DOMAIN.
FT      DOMAIN      151      223      IG-LIKE C2-TYPE DOMAIN.
FT      DISULFID      37      113      POTENTIAL.
FT      DISULFID      158      216      POTENTIAL.
FT      CARBOHYD      70      70      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      137      137      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      173      173      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      186      186      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD      225      225      N-LINKED (GLCNAC...) (POTENTIAL).
FT      VARSPLIC      300      302      GHV -> ESWNLLLLLS (IN ISOFORM 2).
SQ      SEQUENCE      302 AA; 33349 MW; 647934E21B5E34A CRC64;
      MRLGSPGLLF LLFSSLRADT QKEVEKRMVG SDVELSCACP EGSFRDLNDV YVWQTSSEK
      TVYTHIPQN SSSLNVDSRY NRRALMSPAG MLRGDFSLRL FNVTQDEQK FHCILVLSQL
      GFQEVLSVF LNRHVAIVSV PVSAPHSPPS QDELTTCTTS INGYPRPNVY WINKNTNSLL
      DQALQNDTVF LNMRLGYDWW SVLIARTGPS VNIWGCCIENV LLQONLTVGS QTQNDIGERD
      KITEPNVSTG EKNAATVSLT AVCLLVVVA VAIGWVCNDR CLQHSYAGAW AVSPETLTG
      HV
```

SWISS-PROT: O75144

NiceProt - a user-friendly view of this SWISS-PROT entry

ID ICOL HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; Q9HD18;
DT 15-JUL-1999 (Rel. 38, Created)
DT 16-OCT-2001 (Rel. 40, Last sequence update)
DT 01-MAR-2002 (Rel. 41, Last annotation update)
DE ICOS ligand precursor (B7 homolog 2) (B7-H2) (B7-like protein GL50)
DE (B7-related protein-1) (B7RP-1).
GN ICOSL OR B7H2 OR B7RP1 OR KIAA0653.
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. (ISOFORM 1).
RC TISSUE=Dendritic cell;
RX MEDLINE=20477846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
RA Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
RT "Costimulation of T cells by B7-H2, a B7-like molecule that binds
RT ICOS.";
RL Blood 96:2808-2813 (2000).
RN [2]
RP SEQUENCE FROM N.A. (ISOFORM 1), AND CHARACTERIZATION.
RC TISSUE=Peripheral blood lymphocytes;
RX MEDLINE=20465019; PubMed=11007762; [NCBI, ExPASy, EBI, Israel, Japan]
RA Yoshinaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K.,
RA Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H.,
RA Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
RT "Characterization of a new human B7-related protein: B7RP-1 is the
RT ligand to the co-stimulatory protein ICOS.";
RL Int. Immunol. 12:1439-1447 (2000).
RN [3]
RP SEQUENCE FROM N.A. (ISOFORM 2).
RC TISSUE=Leukocyte;
RX MEDLINE=20126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
RA Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A.,
RA Leonard J.P., Hunter S.E., Zoliner R., Thomas J.L., Miyashiro J.S.,
RA Jacobs K.A., Collins M.;
RT "Identification of GL50, a novel B7-like protein that functionally
RT binds to ICOS receptor.";
RL J. Immunol. 164:1653-1657 (2000).
RN [4]
RP SEQUENCE FROM N.A.
RC TISSUE=Brain;
RX MEDLINE=98403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
RA Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,
RA Kotani H., Nomura N., Ohara O.;
RT "Prediction of the coding sequences of unidentified human genes. X.
RT The complete sequences of 100 new cDNA clones from brain which can
RT code for large proteins in vitro.";
RL DNA Res. 5:169-176 (1998).
RN [5]
RP SEQUENCE FROM N.A. (ISOFORM 2).
RA Ling V., Dunussi-Joannopoulos K.;
RT "GL50 molecules and uses thereof.";
RL Patent number WOO121796, 29-MAR-2001.

CC --!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR
CC ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND
CC CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND
CC DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN
CC MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS
CC WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-
CC STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC --!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC --!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2;
CC ARE PRODUCED BY ALTERNATIVE SPLICING.
CC --!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART,
CC KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE
CC MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES,
CC SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN
CC LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC --!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY
CC TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND
CC MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC --!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG
CC SUBFAMILY.
CC --!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
CC --!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
CC --!- CAUTION: Ref.4 sequence differs from that shown in position 300
CC onward for an unknown reason.

CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration
CC between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC the European Bioinformatics Institute. There are no restrictions on its
CC use by non-profit institutions as long as its content is in no way
CC modified and this statement is not removed. Usage by and for commercial
CC entities requires a license agreement (See <http://www.isb-sib.ch/announce/>
CC or send an email to license@isb-sib.ch).
CC -----

DR EMBL; AF199028; AAF34739.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AB014553; BAA31628.1; ALT SEQ. [EMBL / GenBank / DBJ] [CoDingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR MIM; 605717; -. [NCBI / EBI]
DR GeneCards: ICOSL.
DR GeneLynx: ICOSL.
DR Ensembl: O75144.
DR InterPro; IPR003599; Ig.
DR InterPro; IPR003006; Ig MHC.
DR InterPro; IPR003600; Ig like.
DR InterPro; Graphical view of domain structure.
DR Pfam; PF00047; ig; 3.
DR SMART; SM00409; IG; 1.
DR SMART; SM00410; IG like; 1.
DR ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR BLOCKS; O75144.
DR ProtoMap; O75144.
DR PRESAGE; O75144.
DR DIP; O75144.
DR ModBase; O75144.
DR HUGE; KIAA0653; -.
DR SWISS-2DPAGE; GET REGION ON 2D PAGE.

Keywords

KW B-cell activation; Immune response; Glycoprotein;
KW Immunoglobulin domain; Signal; Transmembrane; Multigene family;
KW Alternative splicing.

FT SIGNAL 1 18 POTENTIAL.
FT CHAIN 19 302 ICOS LIGAND.
FT DOMAIN 19 256 EXTRACELLULAR (POTENTIAL).
FT TRANSMEM 257 277 POTENTIAL.
FT DOMAIN 278 302 CYTOPLASMIC (POTENTIAL).
FT DOMAIN 30 120 IG-LIKE V-TYPE DOMAIN.
FT DOMAIN 151 223 IG-LIKE C2-TYPE DOMAIN.
FT DISULFID 37 113 POTENTIAL.
FT DISULFID 158 216 POTENTIAL.
FT CARBOHYD 70 70 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 137 137 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 173 173 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 186 186 N-LINKED (GLCNAC...) (POTENTIAL).
FT CARBOHYD 225 225 N-LINKED (GLCNAC...) (POTENTIAL).
FT VARSPLIC 300 302 GHV -> ESWNLLLLLS (IN ISOFORM 2).
SQ SEQUENCE 302 AA; 33349 MW; 647934E21B5E34A CRC64;
MRLGSPGLLF LLFSSLRADT QKEVEKRMVG SDVELSCACP EGSFRDLNDV YVWQTSSEK
TVYTHIPQN SSSLNVDSRY NRRALMSPAG MLRGDFSRL FNVTQDEQK FHCVLVLSQK
GFQEVLSVF LNRVAANFSV PVSAPHSPPS QDELTTTCIS INGYPRPNVY WINKNTNSLL
DQALQNDTVF LNMRLGYDGV SVLIARTGPS VNIGGCCINVS LLQONLTVGS QTQNDIGERD
KITENPVSTG EKNAATVSLV AIVCLLVVVA VAIGWVCNDR CLQHSYAGAW AVSPETLGT
HV

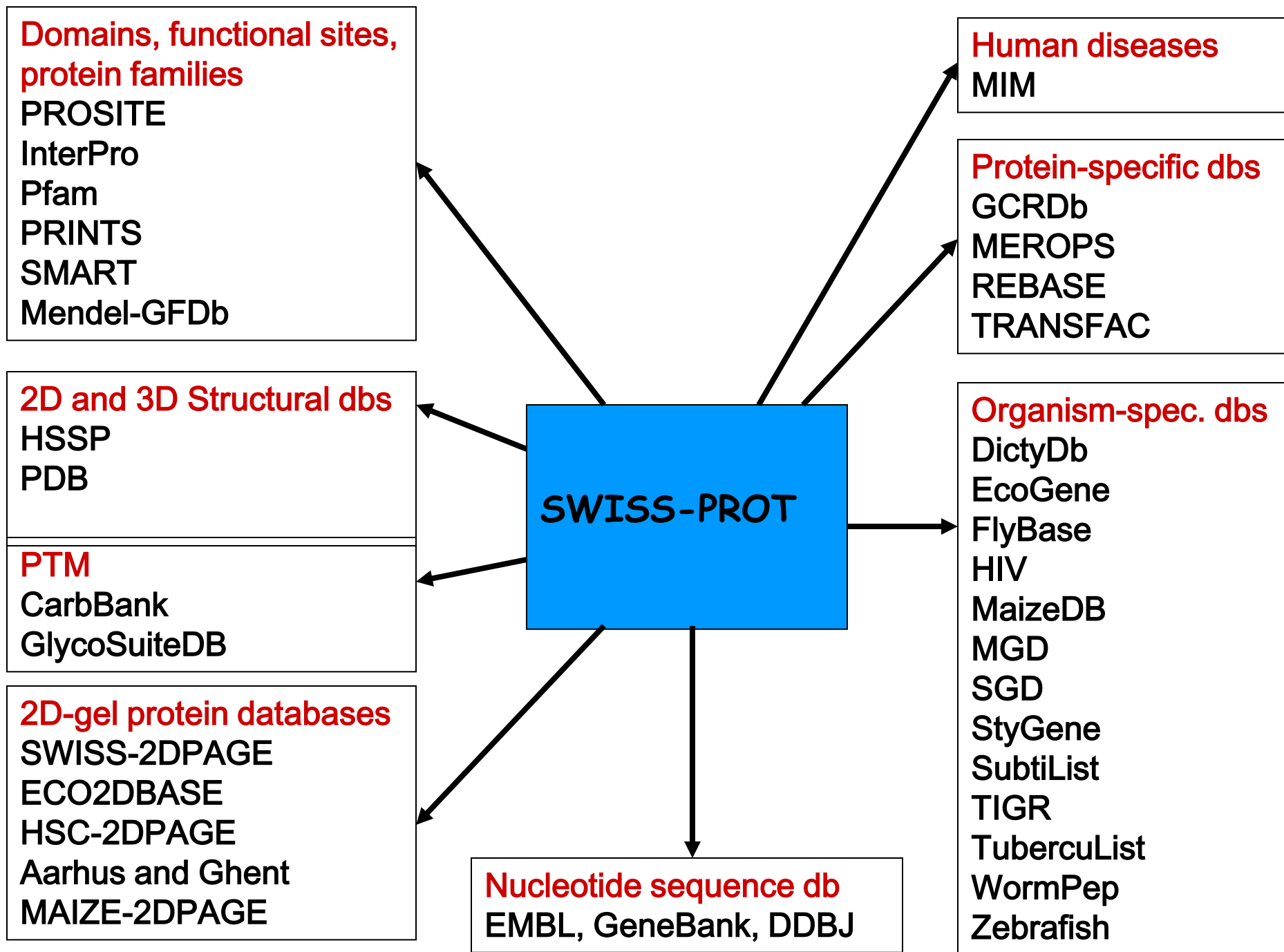
SWISS-PROT: O75144

NiceProt - a user-friendly view of this SWISS-PROT entry

```
ID      ICOL_HUMAN      STANDARD;      PRT;      302 AA.
AC      O75144; Q9NRQ1; Q9HD18;
DT      15-JUL-1999 (Rel. 38, Created)
DT      16-OCT-2001 (Rel. 40, Last sequence update)
DT      01-MAR-2002 (Rel. 41, Last annotation update)
DE      ICOS ligand precursor (B7 homolog 2) (B7-H2) (B7-like protein GL50)
DE      (B7-related protein-1) (B7RP-1).
GN      ICOSL OR B7H2 OR B7RP1 OR KIAA0653.
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. (ISOFORM 1).
RC      TISSUE=Dendritic cell;
RX      MEDLINE=20477846; PubMed=11023515; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.;
RT      "Costimulation of T cells by B7-H2, a B7-like molecule that binds
RT      ICOS.";
RL      Blood 96:2808-2813 (2000).
RN      [2]
RP      SEQUENCE FROM N.A. (ISOFORM 1), AND CHARACTERIZATION.
RC      TISSUE=Peripheral blood lymphocytes;
RX      MEDLINE=20465019; PubMed=11007762; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Yoshinaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K.,
RA      Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H.,
RA      Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.;
RT      "Characterization of a new human B7-related protein: B7RP-1 is the
RT      ligand to the co-stimulatory protein ICOS.";
RL      Int. Immunol. 12:1439-1447 (2000).
RN      [3]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RC      TISSUE=Leukocyte;
RX      MEDLINE=20126021; PubMed=10657606; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A.,
RA      Leonard J.P., Hunter S.E., Zoliner R., Thomas J.L., Miyashiro J.S.,
RA      Jacobs K.A., Collins M.;
RT      "Identification of GL50, a novel B7-like protein that functionally
RT      binds to ICOS receptor.";
RL      J. Immunol. 164:1653-1657 (2000).
RN      [4]
RP      SEQUENCE FROM N.A.
RC      TISSUE=Brain;
RX      MEDLINE=98403880; PubMed=9734811; [NCBI, ExPASy, EBI, Israel, Japan]
RA      Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A.,
RA      Kotani H., Nomura N., Ohara O.;
RT      "Prediction of the coding sequences of unidentified human genes. X.
RT      The complete sequences of 100 new cDNA clones from brain which can
RT      code for large proteins in vitro.";
RL      DNA Res. 5:169-176 (1998).
RN      [5]
RP      SEQUENCE FROM N.A. (ISOFORM 2).
RA      Ling V., Dunussi-Joannopoulos K.;
RT      "GL50 molecules and uses thereof.";
RL      Patent number W00121796, 29-MAR-2001.
```

Feature table (sequence description)

```
CC      -!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR
CC      ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND
CC      CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND
CC      DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN
CC      MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS
CC      WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-
CC      STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC      -!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC      -!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2;
CC      ARE PRODUCED BY ALTERNATIVE SPLICING.
CC      -!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART,
CC      KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE
CC      MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES,
CC      SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN
CC      LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC      -!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY
CC      TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND
CC      MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC      -!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG
CC      SUBFAMILY.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE V-TYPE DOMAIN.
CC      -!- SIMILARITY: CONTAINS 1 IMMUNOGLOBULIN-LIKE C2-TYPE DOMAIN.
CC      -!- CAUTION: Ref.4 sequence differs from that shown in position 300
CC      onward for an unknown reason.
CC      -----
CC      This SWISS-PROT entry is copyright. It is produced through a collaboration
CC      between the Swiss Institute of Bioinformatics and the EMBL outstation -
CC      the European Bioinformatics Institute. There are no restrictions on its
CC      use by non-profit institutions as long as its content is in no way
CC      modified and this statement is not removed. Usage by and for commercial
CC      entities requires a license agreement (See http://www.isb-sib.ch/announce/
CC      or send an email to license@isb-sib.ch).
CC      -----
DR      EMBL; AF199028; AAF34739.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AB014553; BAA31628.1; ALT SEQ. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DBJ] [CoDingSequence]
DR      MIM; 605717; -. [NCBI / EBI]
DR      GeneCards: ICOSL.
DR      GeneLynx: ICOSL.
DR      Ensembl: O75144.
DR      InterPro; IPR003599; Ig.
DR      InterPro; IPR003006; Ig MHC.
DR      InterPro; IPR003600; Ig like.
DR      InterPro; Graphical view of domain structure.
DR      Pfam; PF00047; ig; 3.
DR      SMART; SM00409; IG; 1.
DR      SMART; SM00410; IG like; 1.
DR      ProDom [Domain structure / List of seq. sharing at least 1 domain]
DR      BLOCKS; O75144.
DR      ProtoMap; O75144.
DR      PRESAGE; O75144.
DR      DIP; O75144.
DR      ModBase; O75144.
DR      HUGE; KIAA0653; -.
DR      SWISS-2DPAGE; GET REGION ON 2D PAGE.
KW      B-cell activation; Immune response; Glycoprotein;
KW      Immunoglobulin domain; Signal; Transmembrane; Multigene family;
KW      Alternative splicing.
FT      SIGNAL      1      18      POTENTIAL.
FT      CHAIN       19      302      ICOS LIGAND.
FT      DOMAIN      19      256      EXTRACELLULAR (POTENTIAL).
FT      TRANSMEM    257      277      POTENTIAL.
FT      DOMAIN      278      302      CYTOPLASMIC (POTENTIAL).
FT      DOMAIN      30      120      IG-LIKE V-TYPE DOMAIN.
FT      DOMAIN      151      223      IG-LIKE C2-TYPE DOMAIN.
FT      DISULFID     37      113      POTENTIAL.
FT      DISULFID    158      216      POTENTIAL.
FT      CARBOHYD     70      70      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    137      137      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    173      173      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    186      186      N-LINKED (GLCNAC...) (POTENTIAL).
FT      CARBOHYD    225      225      N-LINKED (GLCNAC...) (POTENTIAL).
FT      VARSPLIC    300      302      GHV -> ESWNLILLLS (IN ISOFORM 2).
SQ      SEQUENCE 302 AA; 33349 MW; 647934E21B5E34A CRC64;
      MRLGSPGLLF LLFSSLRADT QKEVEKRMVG SDVELSCACP EGSFRDLNDV YVWQTSSEK
      TVYTHIPQN SSENVDSDRY NRRALMSFAG MLRGDFSRLR FNVTQDEQK FHCILVLSQK
      GFQEVLSVF LNRVAANFSV PVSAPHSFS QDELTTCTTS INGYPRPNVY WINKNTNSLL
      DQALQNDTVF LNMRLGYDGV SVLIARTPVS VNIGGRIPTV LQQLNLTGVS QTQNDIGERD
      KITEPNVSTG EKNAATVPLG AVLCLLVVVA VAIGWVCNDR CLQHSYAGAW AVSPETLTG
      HV
```



Database 2: Protein sequence

What else ?

Database	Release#	Date	# of Entries
PIR	73.03	16-Aug-2002	283,224
SwissProt	40.26	13-Aug-2002	112,892
TrEMBL	21.7	09-Aug-2002	668,930
GenPept	130.0	15-Jun-2002	1,084,824
RefSeq		21-Aug-2002	379,148
PDB		19-Aug-2002	18,953



PIR

Protein Information Resource

[About PIR](#)[Databases](#)[Search and Retrieval](#)[Download](#)[Support](#)

AN INTEGRATED PUBLIC RESOURCE OF PROTEIN INFORMATICS TO SUPPORT
GENOMIC AND PROTEOMIC RESEARCH AND SCIENTIFIC DISCOVERY

PIR produces the **Protein Sequence Database (PSD)** of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* (1965-1978) edited by Margaret Dayhoff and has been incorporated into an integrated knowledge base system of value-added databases and analytical tools.

[ProClass](#), a central point for exploration of protein information, provides summary descriptions of protein family, function and structure for PIR-PSD, Swiss-Prot, and TrEMBL sequences, with links to over 45 biological databases. [Release 2.3, 12-Aug-2002 contains 809,640 entries](#)

[PIR-NREF](#), a comprehensive database for sequence searching and protein identification, contains non-redundant protein sequences from PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB. [Release 1.04, 26-Aug-2002 contains 979505 entries.](#)



PIR News Flash

View the PIR Informational Flash
New Tools to Databases

Text Search Protein Databases:

GO!

Find an Exact Peptide Match:

GO!

Type in a string of single letter amino acid
code (at least 3 letters)

<http://pir.georgetown.edu/> member databases: PIR-PSD, NRL-3D, ProClass, ProtFam and PIR-ASDB

```

ENTRY      ZUHU #type complete
TITLE      erythropoietin precursor [validated] - human
ORGANISM    #formal name Homo sapiens #common name man
            #cross-references taxon:9606
DATE        27-Nov-1985 #sequence_revision 27-Nov-1985 #text_change
            08-Dec-2000
ACCESSIONS  A01855; A24744; A25384; A22210; S56178
REFERENCE   A01855
            #authors  Jacobs, K.; Shoemaker, C.; Rudersdorf, R.; Neill, S.D.;
            Kaufman, R.J.; Hufson, A.; Seehra, J.; Jones, S.S.;
            Hewick, R.; Fritsch, E.F.; Kavakita, M.; Shimizu, T.;
            Miyake, T.
            #journal    Nature (1985) 313:806-810
            #title      Isolation and characterization of genomic and cDNA clones
            of human erythropoietin.
            #cross-references MUID:85137699
            #accession  A01855
            ##molecule_type mRNA: DNA
            ##residues 1-193 ##label JAC
            ##cross-references GB:X02157; GB:X02158
REFERENCE   A24744
            #authors  Lin, F.K.; Suggs, S.; Lin, C.H.; Browne, J.K.; Smelling,
            R.; Egrie, J.C.; Chen, K.K.; Fox, G.M.; Martin, F.;
            Stabinsky, Z.; Badrawi, S.M.; Lai, P.H.; Goldwasser, E.
            #journal    Proc. Natl. Acad. Sci. U.S.A. (1985) 82:7580-7584
            #title      Cloning and expression of the human erythropoietin gene.
            #cross-references MUID:86067948
            #accession  A24744
            ##molecule_type DNA
            ##residues 1-193 ##label LIN
            ##cross-references GB:M11319; NID:g182197; PIDN:AAA52400.1;
            PID:g182198
REFERENCE   A25384
            #authors  Lai, P.H.; Everett, R.; Wang, F.F.; Arakawa, T.;
            Goldwasser, E.
            #journal    J. Biol. Chem. (1986) 261:3116-3121
            #title      Structural characterization of human erythropoietin.
            #cross-references MUID:86140080
            #accession  A25384
            ##molecule_type protein
            ##residues 28-86,'Q',87-193 ##label LAI
            ##experimental_source urine
            ##note forms without the carboxyl-terminal residue and the four
            carboxyl-terminal residues were observed
REFERENCE   A22210
            #authors  Yanagawa, S.; Hirade, K.; Ohnata, H.; Sasaki, R.; Chiba,
            H.; Ueda, M.; Goto, M.
            #journal    J. Biol. Chem. (1984) 259:2707-2710
            #title      Isolation of human erythropoietin with monoclonal
            antibodies.
            #cross-references MUID:84135751
            #accession  A22210
            ##molecule_type protein
            ##residues 28-29,'X',31-33,'L',35-50,'X',52-53,'D',55,'G',57 ##label
            YAN
REFERENCE   S56178
            #authors  Matsumoto, S.; Ikura, K.; Ueda, M.; Sasaki, R.
            #journal    Plant Mol. Biol. (1995) 27:1163-1172
            #title      Characterization of a human glycoprotein (erythropoietin)
            produced in cultured tobacco cells.
            #cross-references MUID:95284365
            #accession  S56178
            ##molecule_type protein
            ##residues 28-33,'X',35-37 ##label MTS
COMMENT     Erythropoietin is produced by kidney or liver of adult
            mammals and by liver of fetal or neonatal mammals.
GENETICS
            #gene      GDB:EPO
            ##cross-references GDB:119110; OMIM:133170
            #map_position 7q21.3-7q22.1
            #introns    5/1; 53/3; 82/3; 142/3
FUNCTION
            #description the primary inducer of erythrocyte formation
CLASSIFICATION #superfamily erythropoietin
KEYWORDS       erythropoiesis; glycoprotein; hormone; kidney; liver
FEATURE
            1-27      #domain signal sequence #status predicted #label
                     SIG
            28-193    #product erythropoietin #status experimental
                     #label MAT
            34-188,56-60 #disulfide bonds #status experimental\
                     #binding site carbohydrate (Asn) (covalent)
                     #status experimental\
            51,65,110  #binding site carbohydrate (Ser) (covalent)
                     #status experimental\
SUMMARY        #length 193 #molecular_weight 21306
SEQUENCE
            5          10         15         20         25         30
1  MGVHECPA W L W L L S L L S L P L G L P V L G A P P
31 R L I C D S R V L E R Y L L E A K E A E N I T T G C A E H C
61 S L N E N I T V P D T K V N F Y A W K R M E V G Q Q A E V
91 W Q G L A L L S E A V L R G Q A L L V N S S Q P W E P L Q L
121 H V D K A V S G L R S L T T L L R A L G A Q K E A I S P P D
151 A S A A P L R T I T A D T F R K L F R V Y S N F L R G K L
181 K L Y T G E A C R T G D R

```

PIR-PSD: example

« well annotated »

```

COMMENT     Erythropoietin is produced by kidney or liver of adult
            mammals and by liver of fetal or neonatal mammals.

GENETICS
            #gene      GDB:EPO
            ##cross-references GDB:119110; OMIM:133170
            #map_position 7q21.3-7q22.1
            #introns    5/1; 53/3; 82/3; 142/3

FUNCTION
            #description the primary inducer of erythrocyte formation

CLASSIFICATION #superfamily erythropoietin

KEYWORDS       erythropoiesis; glycoprotein; hormone; kidney; liver

FEATURE
            1-27      #domain signal sequence #status predicted #label
                     SIG\
            28-193    #product erythropoietin #status experimental
                     #label MAT\
            34-188,56-60 #disulfide bonds #status experimental\
                     #binding site carbohydrate (Asn) (covalent)
                     #status experimental\
            51,65,110  #binding site carbohydrate (Ser) (covalent)
                     #status experimental\
            153      #binding site carbohydrate (Ser) (covalent)
                     #status experimental

```

2.3 Databases : '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), TAIR(arabidopsis), RGP(rice), MaizeDB (maize), SubtiList (B.subtilis), etc.;
- Generally relational db (Oracle, SyBase or AceDb).

MIM

- 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.



The Arabidopsis Information Resource

TAIR Database ▼

Quick Search

Advanced Search

[Genes](#)

[Markers](#)

[DNA](#)

[Germplasm](#) **NEW**

[Ecotype](#) **NEW**

[Polymorphism/Allele](#)

[People/Labs](#)

[Publications](#)

[Proteins](#)

[Sequences](#)

[GO Annotations](#)

[Locus History](#)

[Microarray](#) **UPDATE**

[More....](#)

Stocks

[About ABRC](#)

[Place Order](#)

[Search Seed Stocks](#)

[Search DNA Stocks](#)

[Search Order History](#)

[Search My Stock Orders](#)

[More....](#)

Analysis Tools

[SeqViewer](#) **UPDATE**

[MapViewer](#)

[AraCyc Pathways](#)

[BLAST](#) **UPDATE**

[WU-BLAST2](#) **UPDATE**

[FASTA](#)

[Patmatch](#)

[Motif Analysis](#) **NEW**

[Bulk Downloads](#)

[Chromosome Map Tool](#) **NEW**

[More....](#)

External Links

[Stock Centers..](#)

[Insertion, Knockout &](#)

[Other Mutations](#)

[Nomenclature](#)

[Sequence Analysis](#)

[Genome Databases](#)

[Proteome Resources](#)

[Microarrays](#)

[More....](#)

Arabidopsis Info

[About Arabidopsis](#)

[Genome Initiative](#)

[Functional Genomics](#)

[Cereon SNPs & Ler](#)

[Gene Expression](#) **NEW**

[Education & Outreach](#)

[Gene Families](#)

[Gene Symbol List](#) **NEW**

[Ontologies](#)

[Data Submission](#)

[Protocols & LabManuals](#)

[More....](#)

FTP Downloads

[Sequences](#)

[Genes](#)

[Maps](#)

[Microarrays](#) **UPDATE**

[Proteins](#)

[Pathways](#) **NEW**

Breaking News

T-DNAs now in SeqViewer

[March 7, 2003]

View T-DNA and transposon insertions in the new SeqViewer T-DNA/Tn band.

Germplasm and Ecotype Searches Released

[March 6, 2003]

Find germplasms using a variety of different criteria that were not available in the old Seed Stock Search.

New positions at TAIR

[March 5, 2003]

Two new positions of a bioinformatics



Search by Name

Search Name:

starts with

(GenBank accessions can also be used)

Restrict by Features

☐ gene structure predicted

☒ has full-length cDNA

☐ is sequenced

☐ is not sequenced

☐ has a clone

☐ is genetic marker

☐ is on a map

Time Restrict

☐ only sequences

Restrict by Map Location

Chromosome

☐ 1

☐ 2

☐ 3

☐ 4

☐ 5

☐ chloroplast genome

☐ mitochondrial genome

Map Type

☐ Physical

☐ Classical

☐ AGI

Range

between

and

Sequence Bulk Download

Paste locus identifiers (such as At1g01030) into the text box. Locus identifiers have to be separated by tabs, commas, carriage returns. Maximum of 1000 ids at a time.

Please use the [ftp server](#) if you want to download entire datasets. This is the sequence datasets used at TAIR.

AGI transcripts

AGI coding sequences

AGI genomic locus sequences

AGI protein sequences

Upstream Sequences - 500bp

Upstream Sequences - 1000bp

Upstream Sequences - 3000bp

Downstream Sequences - 1000bp

Downstream Sequences - 3000bp

Intergenic Sequences

Intron Sequences

Dataset:



2.4 Databases : 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

- [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 \(Finland\)](#)

Mutation/polymorphism: definitions

- **SNPs**: single nucleotide polymorphisms; occur approximately once every 100 to 300 bases.
- **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...
- ! Numbering of the mutated amino acid depends on the db (aa no 1 is not necessary the initiator Met !)

Mutation/polymorphism

The SNP consortium (TSC) <http://snp.cshl.org/>

- Public/private collaboration: Bayer, Roche, IBM, Pfizer, Novartis, Motorola.....
- Has to date discovered and characterized nearly 1.5 million SNPs; in addition, the allele frequencies in three major world populations have been determined on a subset of ~57,000 SNPs.

SNPs dbSNP at NCBI <http://www.ncbi.nlm.nih.gov/SNP/>

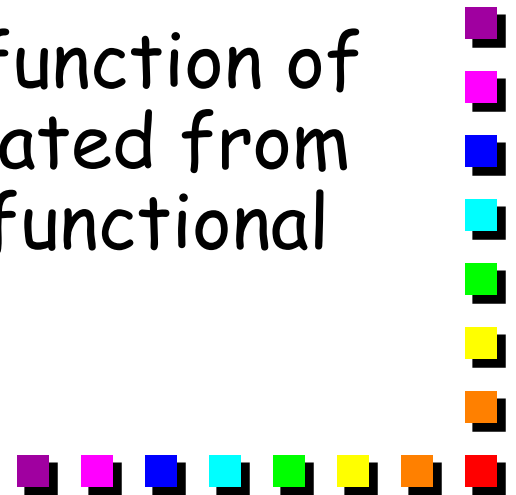
- Collaboration between the National Human Genome Research Institute and the National Center for Biotechnology Information (NCBI)
- Mission: central repository for both single base nucleotide substitutions and short deletion and insertion polymorphisms (several species)
- August 2002, dbSNP has submissions for 4'700'000 SNPs.

Chromosome 21 dbSNP <http://csnp.isb-sib.ch/>

- A joint project between the Division of Medical Genetics of the University of Geneva Medical School and the SIB
- Mission: comprehensive cSNP (Single Nucleotide Polymorphisms within cDNA sequences) database and map of chromosome 21

2.5 Database : protein domain/family

- Contains biologically significant « pattern / profiles/ HMM » formulated in such a way that, with appropriate computational tools, it can rapidly and reliably determine to which known family of proteins (if any) a new sequence belongs to
- -> tools to identify what is the function of uncharacterized proteins translated from genomic or cDNA sequences (« functional diagnostic »)



Protein domain/family

- Most proteins have « modular » structure
- Estimation: ~ 3 domains / protein
- Domains (conserved sequences or structures) are identified by multiple sequence alignments

Sequence ID	start	end	weight	10	20	30	40	50	60
3 <u>EPD HUMAN</u>	2.41			APPRLICDSRVLE	RYLLEAKEAENV	TMGCSEHCSLN	ENITVPD	TKVNFYA	WKRMEVGQQAVEVWQG
2 <u>EPD RAT</u>	2.61			APPRLICDSRVLE	RYILEAKEAENV	TMGCAEGPRL	SENITVPD	TKVNFYA	WKRMEVEEQAIENVQG
3 <u>EPD FELCA</u>	2.99			APPRLICDSRVLE	RYILEAREAEN	ATMGCAEGCSF	SENITVPD	TKVNFYA	WKRMEVGQQALEVWQG
8 Consensus	8.01			APPRLICDSRVLE	RYILEAKEAENV	TMGCAEGCSLN	ENITVPD	TKVNFYA	WKRMEVGQQAVEVWQG
1 PROSITE				-----					

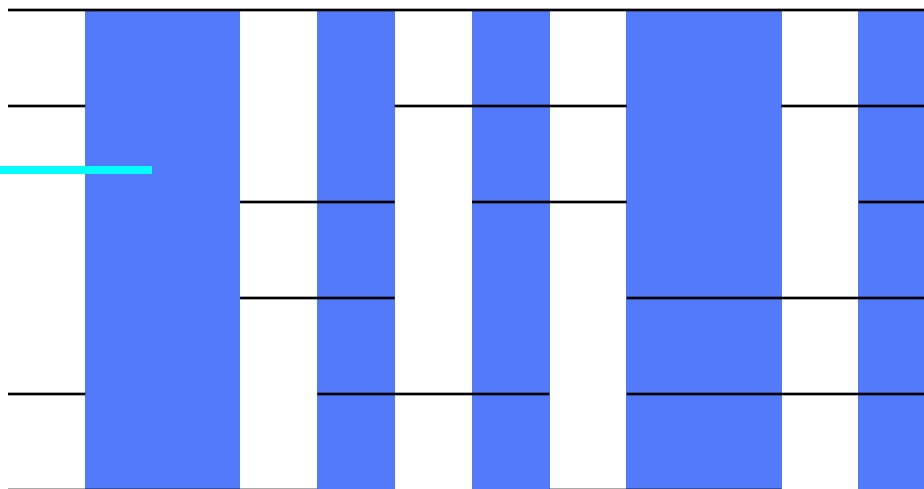
- Domains can be defined by different methods:
 - Pattern (regular expression); used for very conserved domains
 - Consensus sequence
 - Profiles (weighted matrices): two-dimensional tables of position specific match-, gap-, and insertion-scores, derived from aligned sequence families; used for less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

Methodology

Sequence ID	start	end	weight	10	20	30	40	50	60
3 ⊕ EPO HUMAN	2.41			APPRLICDSRVLEPYLLEAKEAENV	TMGCSEHCSL	NENITVPD	TKVNFYAWKRMEV	GQQAVEV	WQG
2 ⊕ EPO RAT	2.61			APPRLICDSRVLEPYILEAKEAENV	TMGCAEGPRL	SENITVPD	TKVNFYAWKRMEV	EEQAIEV	WQG
3 ⊕ EPO FELCA	2.99			APPRLICDSRVLEPYILEAREAENAT	MGCAEGCSF	SENITVPD	TKVNFYAWKRMEV	QQQALEV	WQG
8 Consensus	8.01			APPRLICDSRVLEPYILEAKEAENV	TMGCAEGCSL	NENITVPD	TKVNFYAWKRMEV	GQQAVEV	WQG
1 PROSITE									

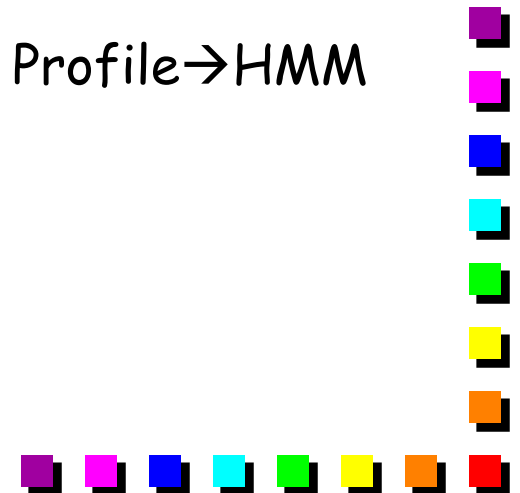
cydeggis
cyedggis
cyeeddit
cyngdggs
cyrgdgnt

c-y-x2-[dg]-g-x-[st]

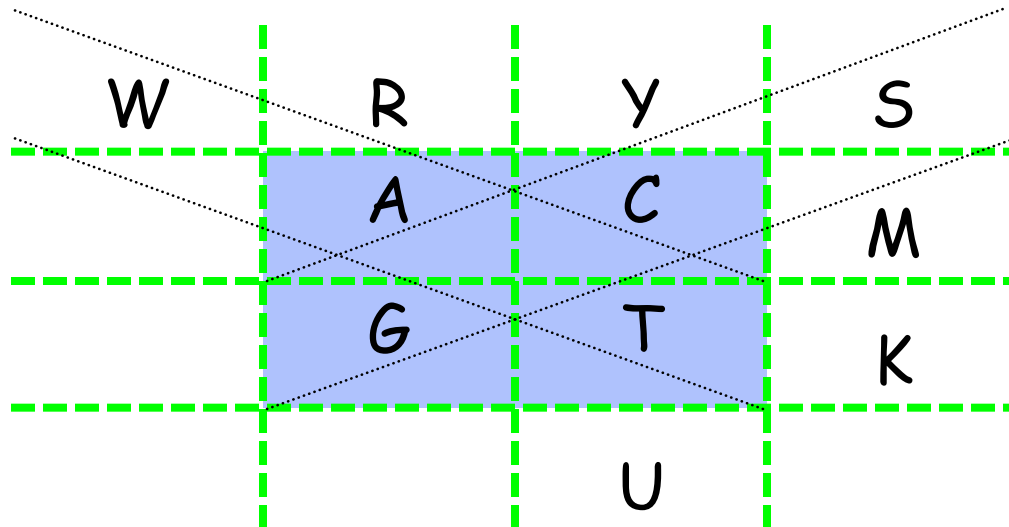


Regular expression/ weighted matrices

Profile → HMM



Consensus sequence



B: not A

D: not C

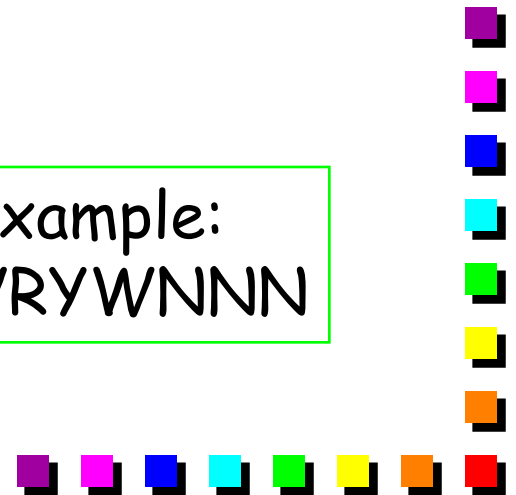
H: not G

V: not T

N: any

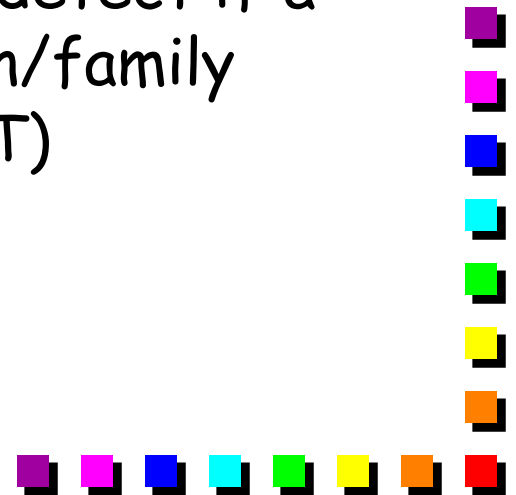
X: unknown

A example:
GTRWRYWNNN



Protein domain/family db

- **Secondary databases** are the fruit of analyses of the sequences found in the primary sequence db
- Either manually curated (i.e. PROSITE, Pfam, etc.) or automatically generated (i.e. ProDom, DOMO)
- Some depend on the method used to detect if a protein belongs to a particular domain/family (patterns, profiles, HMM, PSI-BLAST)



Protein domain/family db

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

I
n
t
e
r
p
r
o

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





History and numbers

- Founded by Amos Bairoch
- 1988 First release in the PC/Gene software
- 1990 Synchronisation with Swiss-Prot
- 1994 Integration of « profiles »
- 1999 PROSITE joins InterPro
- August 2002 Current release 17.19
 - 1148 documentation entries
 - 1568 different patterns, rules and profiles/matrices with list of matches to SWISS-PROT



Prosite (pattern): example

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: 40.7, total number of sequence entries in that release: 103373. • 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 %	



Prosite (pattern): example

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**

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 (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):
	TPO_PIG (P42706)
	Retrieve an alignment of SWISS-PROT true positive hits:
	[Clustal format, color, condensed view] [Clustal format, color] [Clustal format, plain text] [Fasta format]



Prosite (profile): example

General information about the entry	
Entry name	INTEIN_C_TER
Accession number	PS50818
Entry type	MATRIX
Date	MAY-2002 (CREATED); MAY-2002 (DATA UPDATE); MAY-2002 (INFO UPDATE).
PROSITE documentation	PDOC00687
Name and characterization of the entry	
Description	Intein C-terminal splicing motif profile.
Matrix / Profile	<pre> /GENERAL_SPEC: ALPHABET='ABCDEFGHIJKLMNPQRSTVWYZ'; LENGTH=22; /DISJOINT: DEFINITION=PROTECT; N1=3; N2=20; /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.8533; R2=0.02263959; TEXT='NScore'; /CUT_OFF: LEVEL=0; SCORE=290; N_SCORE=7.4; MODE=1; TEXT='!'; /CUT_OFF: LEVEL=-1; SCORE=249; N_SCORE=6.5; MODE=1; TEXT='?'; /DEFAULT: M0=-8; D=-20; I=-20; B0=-60; B1=-60; E0=-60; E1=-60; MI=-105; MD=-105; IM=-105; DM=-105; A B C D E F G H I K L M N P Q R S T V W Y Z /I: B0=0; B1=0; BI=-105; BD=-105; /M: SY='Y'; M=-15,-11,-29,-9,-6, 3,-23,-5,-14,-1,-13,-10,-10,-2,-9, 0,-12,-8,-15,-5,12,-9; /M: SY='V'; M= 1,-25,-13,-27,-26,-1,-24,-27,20,-18, 7, 6,-24,-26,-25,-17,-8, 1,34,-27,-9,-26; /M: SY='Y'; M=-19,-17,-28,-18,-17,29,-28, 9,-2,-11,-1,-1,-16,-27,-11,-9,-18,-10,-9,17,55,-17; /M: SY='D'; M=-13,27,-1,33, 5,-29,-12,-6,-30,-8,-25,-23,15,-17,-6,-12, 4,-3,-22,-41,-21,-1; /M: SY='L'; M=-6,-25,-20,-27,-22, 3,-30,-24,23,-25,26,13,-24,-25,-21,-21,-18,-4,20,-23,-3,-23; /M: SY='T'; M=-1, 1,-18,-1, 4,-16,-10,-10,-14,-7,-16,-13, 3,-12,-3,-9,11,12,-9,-28,-11, 0; /M: SY='V'; M= 3,-18,-3,-21,-22,-8,-21,-25,10,-18, 1, 0,-17,-24,-22,-19,-4, 4,23,-30,-13,-22; /I: I=-5; MD=-27; /M: SY='E'; M=-10, 0,-25, 7,15,-17,-19,-7,-21, 2,-19,-15,-5,11, 0,-6,-6,-9,-21,-21,-9, 6; D=-5; /I: I=-5; MD=-27; /M: SY='N'; M=-2,10,-17, 7, 0,-18, 5,-1,-18,-2,-18,-12,13,-11,-3,-3, 2,-3,-16,-22,-13,-2; D=-5; /I: I=-5; MI=-27; MD=-27; IM=-27; DM=-27; /M: SY='H'; M=-12, 0,-26, 1, 5,-22,-6,42,-25,-8,-18,-6, 5,-16, 7,-4,-4,-13,-24,-27, 0, 4; D=-5; /I: I=-5; DM=-27; /M: SY='N'; M=-5,14,-19, 2,-5,-10,-8,-2,-13,-4,-15,-11,27,-20,-5,-2, 3, 0,-17,-30,-13,-5; /M: SY='F'; M=-16,-25,-24,-30,-25,40,-25,-8, 5,-22, 9, 3,-20,-29,-25,-17,-20,-10, 0, 8,37,-25; /I: I=-6; MD=-32; /M: SY='V'; M=-6,-27,-18,-30,-25, 7,-30,-20,24,-24,18,12,-23,-26,-23,-20,-16,-4,25,-21,-1,-25; D=-6; /I: I=-6; MI=-32; IM=-32; DM=-32; /M: SY='A'; M=15,-14,-14,-20,-15,-7,-14,-20, 1,-15,-2,-3,-12,-17,-13,-17, 4, 8, 8,-23,-11,-14; /I: I=-5; MD=-25; </pre>



Prosite (profile): example

Numerical results

- SWISS-PROT release number: **40.17**, total number of sequence entries in that release: **108489**.
- Total number of hits in SWISS-PROT: **68 hits in 59 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **68 hits in 59 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: **0**
- Precision (true hits / (true hits + false positives)): **100.00 %**
- Recall (true hits / (true hits + false negatives)): **100.00 %**

Comments

- MATRIX_TYPE: **protein_domain**
- Scaling database: **SWISS-PROT: reversed**
- Matrix author: **CJA_Sigrist**
- Taxonomic range: **Archaeobacteria, Eukaryotes, Prokaryotes (Bacteria), Eukaryotic viruses**
- Maximum known number of repetitions of the pattern in a single protein: **3**

Cross-references

True positive hits:

DNAB_GUITH ([Q78411](#)), DNAB_MYCTU ([P71715](#)), DNAB_PORPU ([P51333](#)),
DNAB_MYCTU ([Q58455](#)), DNAB_GUITH ([Q55118](#)), YB20_PORPU ([Q58455](#)),
Y832_METJA ([Q58242](#)), YA54_METJA ([Q58454](#)), YE20_METJA ([Q58815](#)),
YE61_MYCLE ([Q49689](#)), YE61_MYCTU ([Q53152](#))

Retrieve an alignment of SWISS-PROT true positive hits:

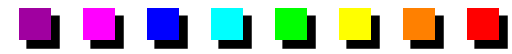
[\[Clustal format, color, condensed view\]](#) [\[Clustal format, color\]](#) [\[Clustal format, plain text\]](#) [\[Fasta format\]](#)

PDB

new

[1AM2](#);

[\[Detailed view\]](#)



BLOCKS database

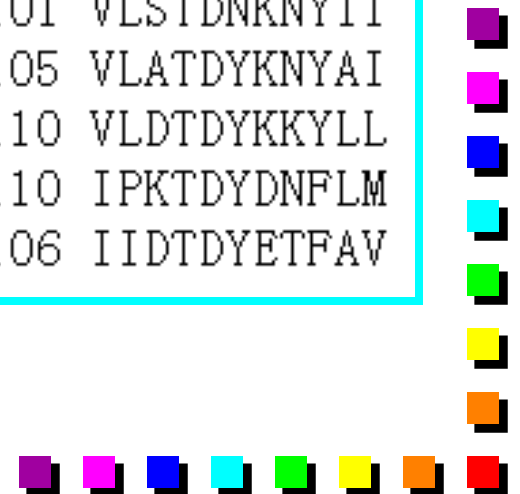
- Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins

```
>Lipocal Five tough ones from Lawrence et al, Science 262:  
5 sequences are included in 2 blocks
```

	LipocalA, width = 15	LipocalB, width = 11
BBP_PIEBR	16 NFDWSNYHGKWWWEVA (70)	101 VLSTDNKNYII
ICYA_MANSE	17 DFDLSAFAGAWHEIA (73)	105 VLATDYKNYAI
LACB_BOVIN	25 GLDIQKVAGTWYSLA (70)	110 VLDTDYKKYLL
MUP2_MOUSE	27 NFNVEKINGEWHTII (68)	110 IPKTDYDNFLM
RETB_BOVIN	14 NFDKARFAGTWYAMA (77)	106 IIDTDYETFAV

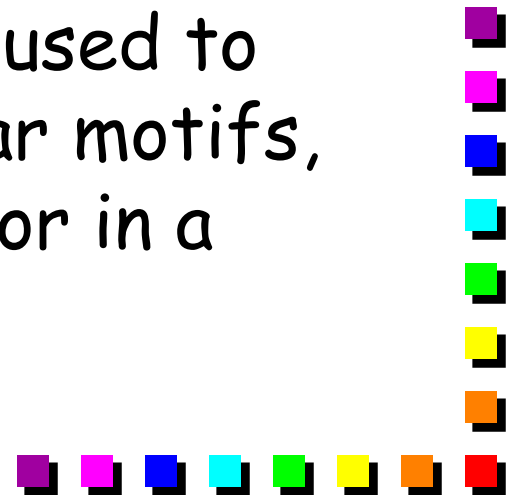
↑
positions

↑
weights



PRINTS database

- PRINTS is a compendium of protein **fingerprints**. A fingerprint is a group of conserved motifs used to characterize a protein family
- Usually the motifs do not overlap, but are separated along a sequence
- A fingerprint is a set of motifs used to predict the occurrence of similar motifs, either in an individual sequence or in a database.



Pfam Protein families database of alignments

Home Keyword Search Protein Search Browse Pfam DNA Search Taxonomy Rf GP120

Accession number: PF00061

Envelope glycoprotein GP120

The entry of HIV requires interaction of viral GP120 with [CD4](#), [CXCR4](#) and a coreceptor. This family forms structural complexes with other Pfam families, to view the description (entry) click on the [HMM logo](#).

The entry of HIV requires interaction of viral GP120, an envelope glycoprotein, with the CD4 receptor on the cell surface. Proteins belonging to the family are found in the [HMM logo](#).

Figure 1: 1ce4
Amphipathic helix
Conformational model for the conserved C2 loop of the envelope protein gp120 of HIV-1

Alignment

Seed (24) Full (34964)

Format Coloured alignment

Get alignment

Domain on

Seed (24)

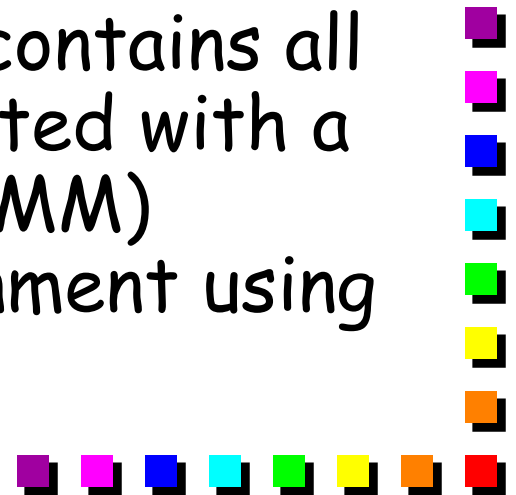
As a Graphic

Zoom 0.5 pixels/aa

View Graphic

Pfam database

- Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families
- Each curated family in Pfam is represented by a seed and full alignment. The seed contains representative members of the family, while the full alignment contains all members of the family as detected with a profile hidden Markov model (HMM) constructed from the seed alignment using the HMMER2 software



InterPro is an integrated documentation resource for protein families, domains and sites.

- Sequence-motif methods:
 - PROSITE, home of regular expressions and profiles;
 - Pfam, SMART and TIGRFAMs, keepers of hidden markov models (HMMs);
 - PRINTS, provider of fingerprints (groups of aligned, un-weighted motifs);
- Diagnostically, these resources have different areas of optimum application owing to the different underlying analysis methods. In terms of family coverage, the protein signature databases are similar in size but differ in content.
- While all of the methods share a common interest in protein sequence classification, some focus on divergent domains (e.g., Pfam), some focus on functional sites (e.g., PROSITE), and others focus on families, specialising in hierarchical definitions from superfamily down to subfamily levels in order to pin-point specific functions (e.g., PRINTS). TIGRFAMs focus on building HMMs for functionally equivalent proteins.



InterPro release 6.0 (2003.3)

- Contains 7751 entries, representing 1748 domains, 5841 families, 147 repeats, and 15 post-translational modification sites. Member databases:

Database		Version	Entries
SWISS-PROT	→ Primary databases	41.0	122564
PRINTS		35.0	1750
TrEMBL	→	23.0	830525
Pfam		8.0	5193
PROSITE patterns		17.37	1605
PROSITE preprofiles		N/A	150
ProDom		2002.1	1021
InterPro		6.0	7751
Smart		3.4	654
TIGRFAMs		2.1	1614



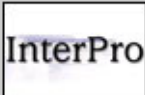
InterPro: www.ebi.ac.uk/interpro

**EMBL-EBI**
European Bioinformatics Institute

Get for ? Site search

EBI HomeAbout EBIResearchServicesToolboxDatabasesDownloadsSubmissions

INTERPRO DATABASE

**InterPro**

- [InterPro Index](#)
- [Text Search](#)
- [Sequence Search](#)
- [Databases](#)
- [Documentation](#)
- [FTP Site](#)

InterPro

InterPro is a useful resource for whole genome analysis and has already been used for the [proteome analysis](#) of a number of completely sequenced organisms including *preliminary* analyses of the mouse and human genomes.



Further information on InterPro can be found in the [Documentation](#) page, which includes links to the [release notes](#), the [user manual](#), [a list of deleted InterPro entries](#), the [dataflow scheme](#) of the database, a fully annotated [sample entry](#) and [references](#) for the [member databases](#).

InterPro is headed by **Rolf Apweiler**.

Updated Documents and New Links

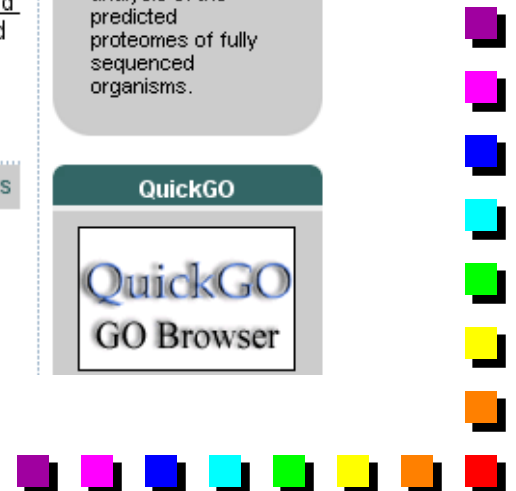
- **Announcement:** **InterPro release 5.1** is out with new data and updated files.
- **News:** InterPro has a new SRS-based text search which allows users to search a combination of InterPro and protein features.
- List of all InterPro entries of each type

Proteome Analysis

Statistical and comparative analysis of the predicted proteomes of fully sequenced organisms.

QuickGO

QuickGO
GO Browser



InterPro example

InterPro Entry IPR001323

Erythropoietin/thrombopoietin

Database	InterPro
Accession	IPR001323; EPO_TPO (matches 21 proteins)
Name	Erythropoietin/thrombopoietin
Type	Family 
Dates	08-OCT-1999 (created) 23-NOV-2000 (last modified)
Signatures	PS00817; EPO_TPO (19 proteins) PF00758; EPO_TPO (21 proteins)
Children  [tree]	IPR003013; Erythropoietin (12 proteins) IPR003978; Thrombopoietin (5 proteins)
Function 	glycopeptide hormone (GO:0005181)
Component 	extracellular (GO:0005576)
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 <u>mammals</u>, i.e. that from one species, say <u>human</u>, can stimulate erythropoiesis in other species, say <u>mouse</u> or <u>rat</u> [4]. Thrombopoietin (TPO), a glycoprotein, is the <u>mammalian</u> 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.



InterPro example

Examples	• P49745 TPO_RAT • P33709 EPO_SHEEP • P33708 EPO_FELCA View examples
References	1. Shoemaker C.B., Mitscock L.D. <i>Murine erythropoietin gene - Cloning, expression, and human gene homology.</i> Mol. Cell. Biol. 6: 849-858(1986). [MEDLINE:87039105] 2. Takeuchi M., Takasaki S., Miyazaki H., Kato T., Hoshi S., Kochibe N., Kobata A. <i>Comparative study of the asparagine-linked sugar chains of human erythropoietins purified from urine and the culture medium of recombinant chinese hamster ovary cell.</i> J. Biol. Chem. 263: 3657-3663(1988). [MEDLINE:88153657] 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. <i>Monkey erythropoietin gene - Cloning, expression and comparison with the human erythropoietin gene.</i> Gene 44: 201-209(1986). [MEDLINE:87055236] 4. Nagao M., Suga H., Okano M., Masuda S., Narita H., Ikura K., Sasaki R. <i>Nucleotide sequence of rat erythropoietin.</i> Biochim. Biophys. Acta 1171: 99-102(1992). [MEDLINE:93042015]
Database links	PROSITE doc; PDOC00644 Blocks; IPB001323
Matches 	Table all Graphical all Condensed graphical view



InterPro graphic example

InterPro - Proteins matching IPR001323

Table Graphical



Grid shows 10aa intervals, first mark at position 0. Move the mouse over a match to see more information in the status line of your browser window.

Item 1-20 of 21

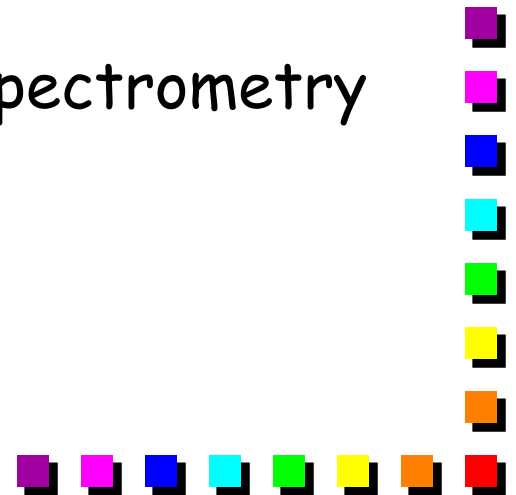
< 1 2 >

Protein	Match Display
SWISS-PROT EPO_HUMAN P01588	IPR001323 PS00817  EPO_TPO
	IPR001323 PF00758  EPO_TPO
	IPR003013 PR00272  ERYTHROPTN
SWISS-PROT EPO_MOUSE P07321	IPR001323 PS00817  EPO_TPO
	IPR001323 PF00758  EPO_TPO
	IPR003013 PR00272  ERYTHROPTN
SWISS-PROT EPO_MACFA P07865	IPR001323 PS00817  EPO_TPO
	IPR001323 PF00758  EPO_TPO
	IPR003013 PR00272  ERYTHROPTN
SWISS-PROT EPO_RAT P29676	IPR001323 PS00817  EPO_TPO
	IPR001323 PF00758  EPO_TPO
	IPR003013 PR00272  ERYTHROPTN
SWISS-PROT EPO_CANFA P33707	IPR001323 PS00817  EPO_TPO
	IPR001323 PF00758  EPO_TPO

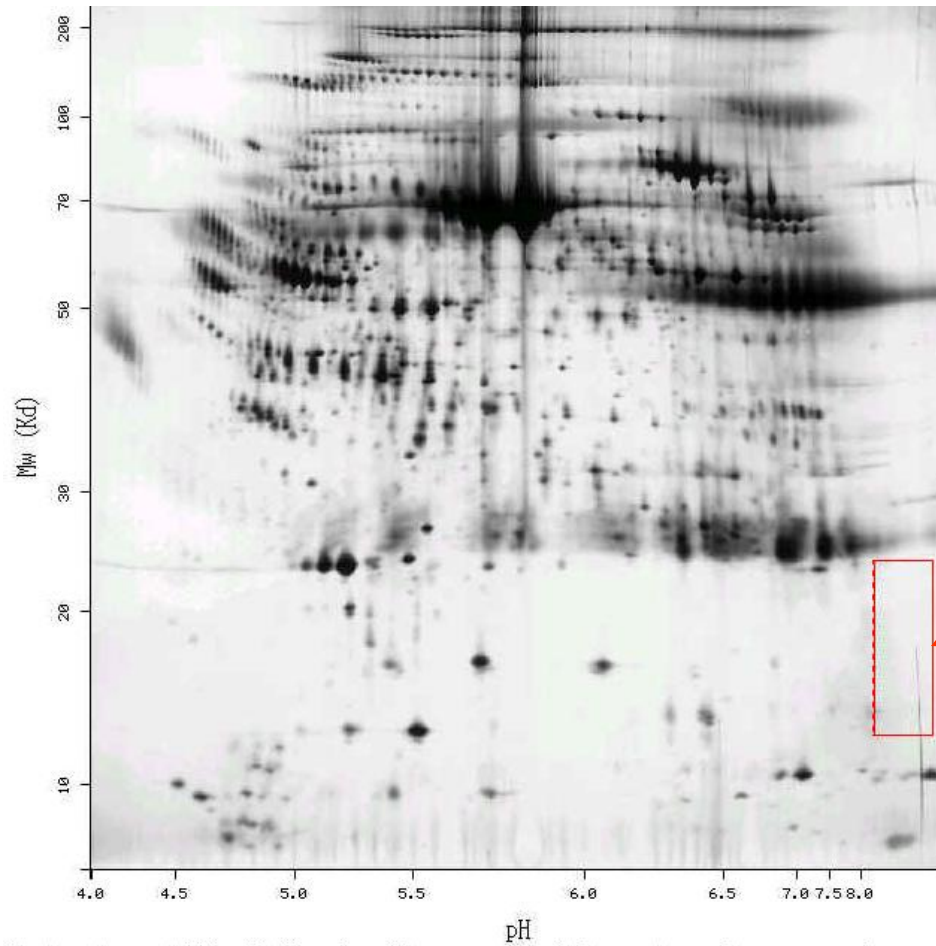


2.6 Databases : proteomics

- Contain informations obtained by 2D-PAGE: master images of the gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Format: composed of image and text files
- Most 2D-PAGE databases are "federated" and use SWISS-PROT as a master index
- There is currently no protein Mass Spectrometry (MS) database (not for long...)



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

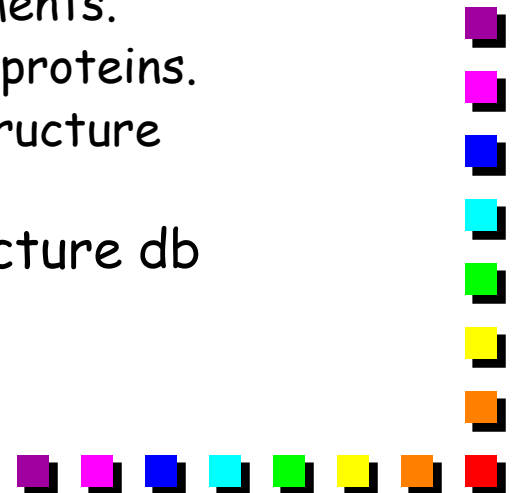


EPO_HUMAN
(human plasma)
Should be here...



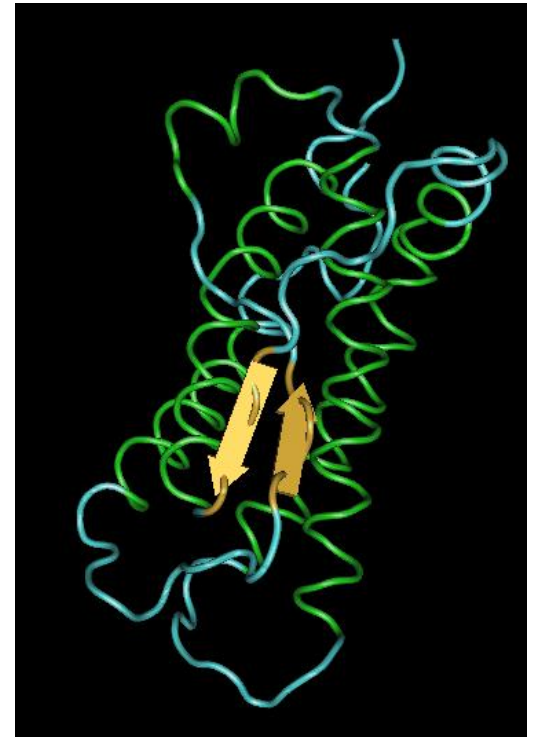
2.7 Databases : 3D structure

- 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, virus, complex protein/DNA...)
- RCSB or PDB (Protein Data Bank), CATH and SCOP (structural classification of proteins (according to the secondary structures)), BMRB (BioMagResBank; NMR results)
- DSSP: Database of Secondary Structure Assignments.
HSSP: Homology-derived secondary structure of proteins.
FSSP: Fold Classification based on Structure-Structure Assignments.
- SWISS-MODEL: Homology-derived 3D structure db



RCSB or PDB: Protein Data Bank

- Managed by Research Collaboratory for Structural Bioinformatics (RCSB) (USA).
- Contains macromolecular structure data on proteins, nucleic acids, protein-nucleic acid complexes, and viruses.
- Specialized programs allow the visualization of the corresponding 3D structure. (e.g., SwissPDB-viewer, Cn3D)
- Currently there are ~18'000 structure data for 6'000 different molecules, but far less protein family (highly redundant) !



EPO_HUMAN



Sequence and secondary structure

5'-Nucleotidase From E. Coli

```

1  MKLLQRGVAL  LLTETETL  LG  ETAL  LNEQRY  EYHLETH  HTL  EYHGHHEHNE
   T

51  YGEYGLAAQF  PD
   TS  HHHHHH  HH

101  FRGMNLVGYI  ER
   HHHHHHHT  EE

151  LFKPWALFKI  VI
   SSBSEEEEEH  HH

201  QELQQTEKPI  VG
   HHHHHHHT  E

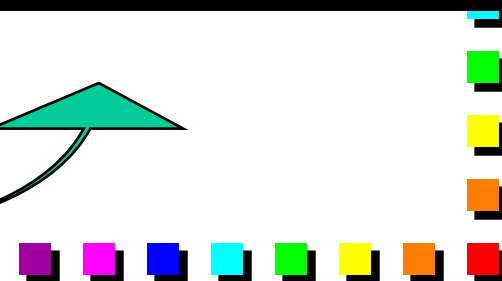
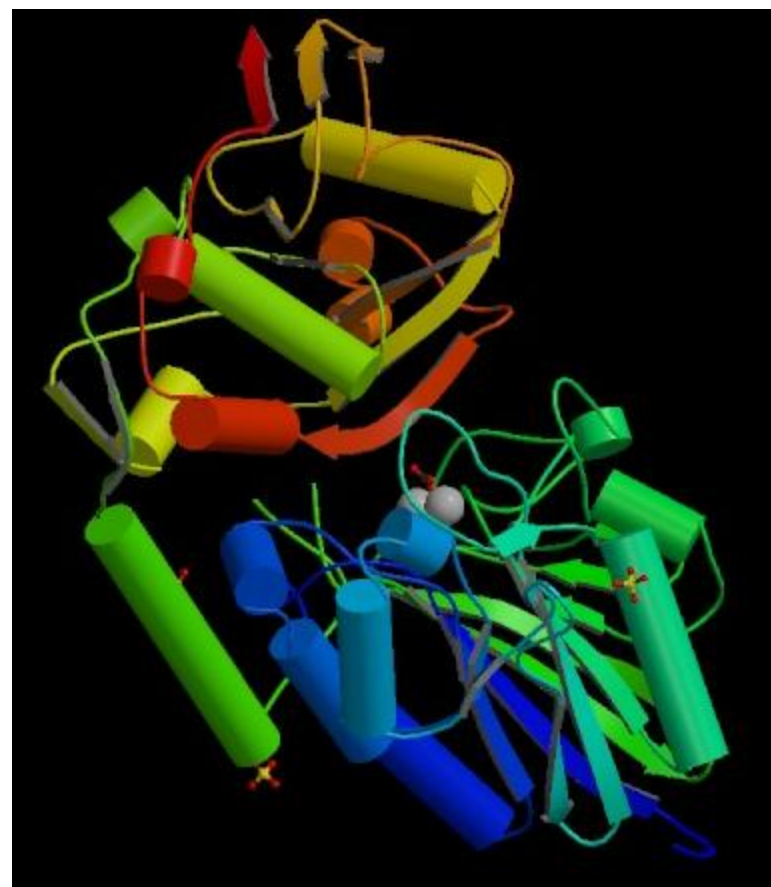
251  GHSQDPVCM  DF
   SS  B  BS  EE

301  EFRNGEMKM  LS
   EEETEEEEH  HH

351  PFQNKGKAQI  AD
   THHHHHHHHT  S

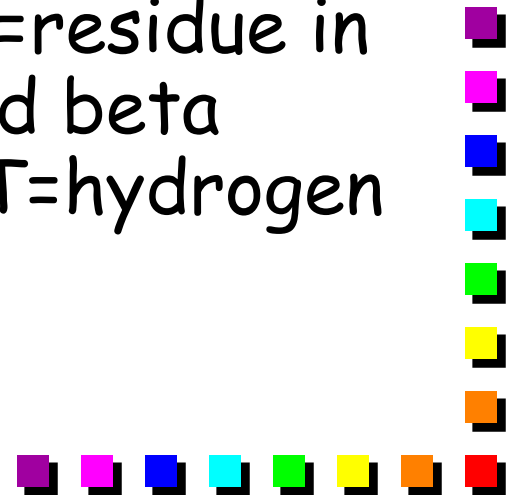
401  FAVMSGGGIR  DSIKAGDISY  KNVLKVQPF  G  NVVVYADMIG  KEVIDYLTAV
   EEEEEGGG  B  S EEEEEH  HHHHHHS  S  EEEEEEEH  HHHHHHHHHH

451  AQMKPDSGAY  PQFANVSFVA  KDGKLNLDKI  KGEPVDPAKT  YRMATLNFNA
   TTS SSSTT  EEEEEEEEE  ETTEES  EE  TTEE  TTSE  EEEEEHHHH
  
```



PDB: Sequence Details

- Each residue in the sequence is reported as a single letter code. Secondary structure is calculated and described according to an implementation of the method of Kabsch and Sander (1983) *Biopolymers* 22, 2577-2637.
- The assignments are: H=helix; B=residue in isolated beta bridge; E=extended beta strand; G=310 helix; I=pi helix; T=hydrogen bonded turn; S=bend.

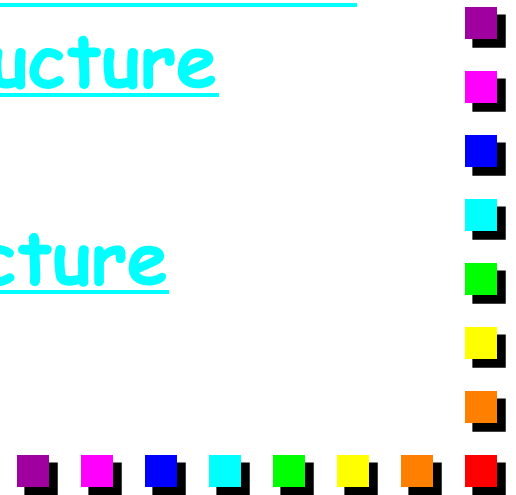


	000000	0.007326	0.000000	
1	-38.912	14.988	99.206	1.00 74.25
1	-37.691	14.156	98.995	1.00 72.12
1	-36.476	15.045	98.733	1.00 70.30
1	-36.607	16.130	98.160	1.00 68.80
1	-37.910	13.201	97.819	1.00 70.67
2	-35.278	14.597	99.162	1.00 70.55
2	-34.022	15.337	98.982	1.00 66.55

SHEET	2	I 4 ILE C 154 ALA C 162 -1 N VAL C 158 O VAL C 172
SHEET	3	I 4 ARG C 191 MET C 200 -1 N ARG C 199 O ARG C 155
SHEET	4	I 4 VAL C 216 LEUC C 219 -1 N LEUC C 218 O TYR C 192
SSBOND	1	CYS A 7 CYS A 161
SSBOND	2	CYS A 29 CYS A 33
SSBOND	3	CYS B 28 CYS B 38
SSBOND	4	CYS B 67 CYS B 83
SSBOND	5	CYS C 28 CYS C 38
SSBOND	6	CYS C 67 CYS C 83
CISPEP	1	GLU B 202 PRO B 203 0 0.05
CISPEP	2	GLU C 202 PRO C 203 0 0.14
CRYST1	58.400 79.300 136.500 90.00 90.00 90.00 P 21 21 21 4	
ORIGX1	1.000000 0.000000 0.000000 0.000000	
ORIGX2	0.000000 1.000000 0.000000 0.000000	
ORIGX3	0.000000 0.000000 1.000000 0.000000	
SCALE1	0.017123 0.000000 0.000000 0.000000	
SCALE2	0.000000 0.012610 0.000000 0.000000	
SCALE3	0.000000 0.000000 0.007326 0.000000	
ATOM	1 N ALA A 1 -38.912 14.988 99.206 1.00 74.25	
ATOM	2 CA ALA A 1 -37.691 14.156 98.995 1.00 72.12	
ATOM	3 C ALA A 1 -36.476 15.045 98.733 1.00 70.30	
ATOM	4 O ALA A 1 -36.607 16.130 98.160 1.00 68.80	
ATOM	5 CB ALA A 1 -37.910 13.201 97.819 1.00 70.67	
ATOM	6 N PRO A 2 -35.278 14.597 99.162 1.00 70.55	
ATOM	7 CA PRO A 2 -34.022 15.337 98.982 1.00 66.55	
//////		

Secondary structural databases

- SCOP : Structural Classification of Protein
- CATH : Structural Classification
(class/architecture/topology/homology)
- FSSP : Fold classification based on
Structure-Structure alignment of Proteins
- MMDB: a Molecular Modeling Database in
NCBI's macromolecular 3D Structure
database
- PDBsum: Summary of PDB Structure

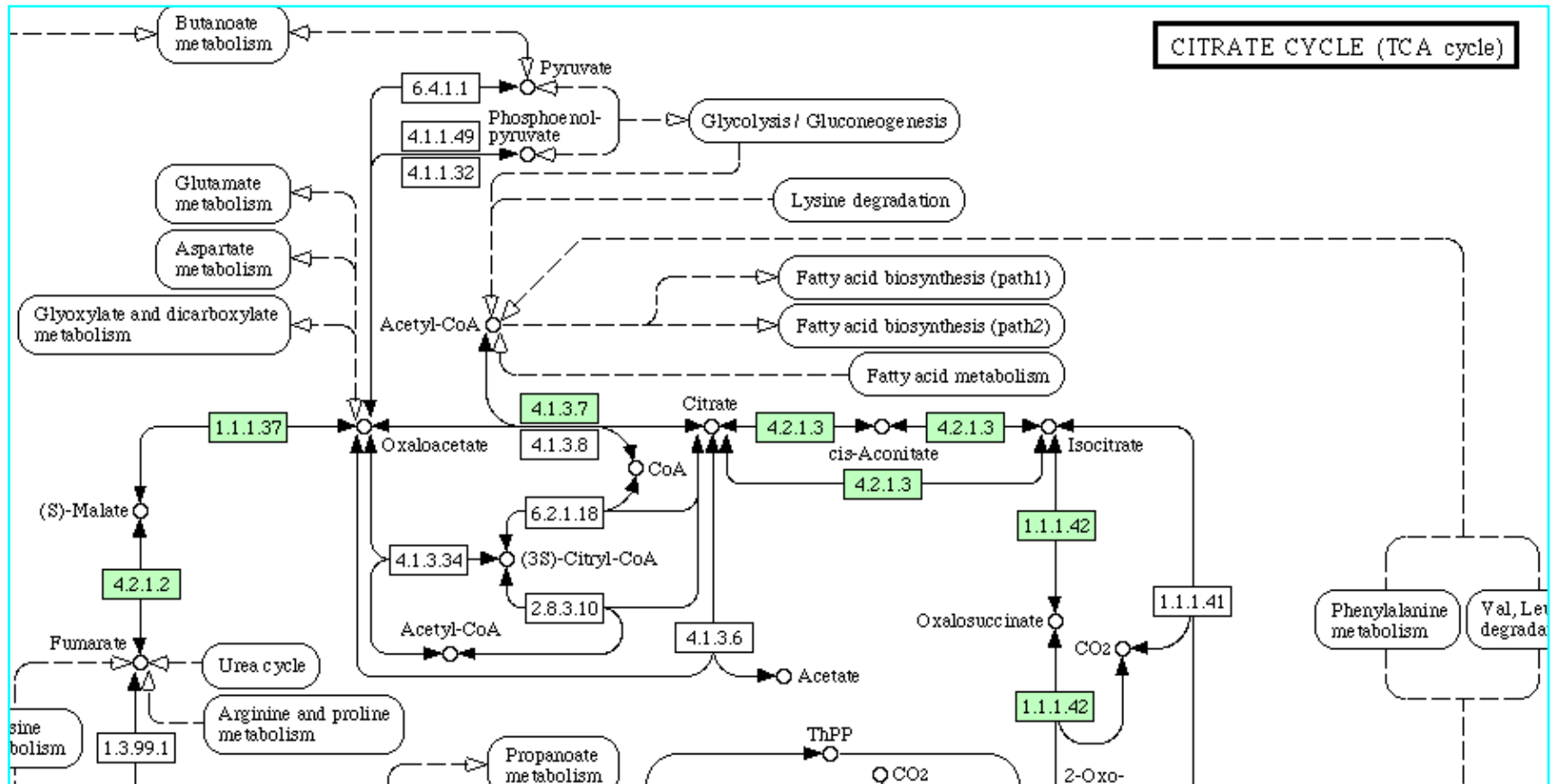


2.8 Databases : metabolic

- 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.



KEGG: Kyoto Encyclopedia of Genes and Genomes



www.genome.ad.jp/kegg/kegg2.html



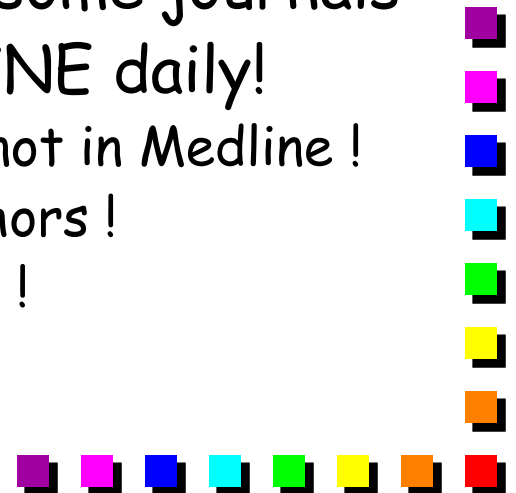
2.9 Databases : bibliographic

- Bibliographic reference databases contain citations and abstract informations of published life science articles;
- Example: Medline
- Other more specialized databases also exist (example: Agricola).



Medline

- MEDLINE covers the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences
- more than 4,600 biomedical journals published in the United States and 70 other countries
- Contains over 11 million citations since 1966 until now
- Contains links to biological db and to some journals
- New records are added to PreMEDLINE daily!
 - Many papers not dealing with human are not in Medline !
 - Before 1970, keeps only the first 10 authors !
 - Not all journals have citations since 1966 !



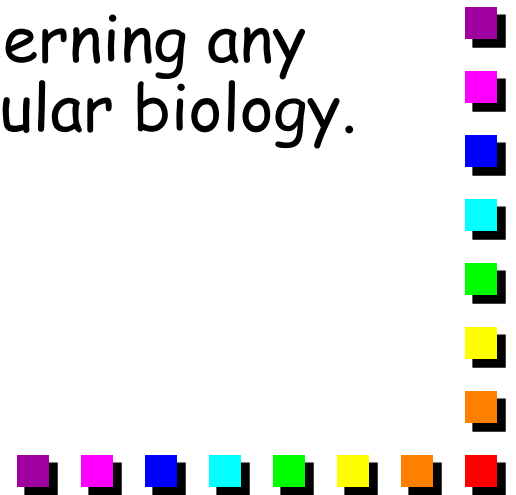
Medline/Pubmed

- PubMed is developed by the National Center for Biotechnology Information (NCBI)
- PubMed provides access to bibliographic information such as MEDLINE, PreMEDLINE, HealthSTAR, and to integrated molecular biology databases (composite db)
 - PMID: 10923642 (PubMed ID)
 - UI: 20378145 (Medline ID)



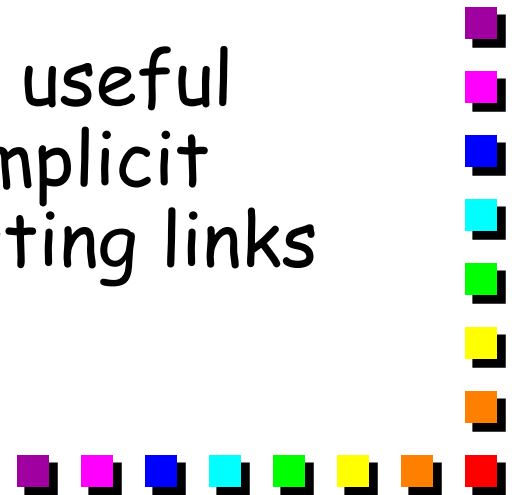
2.10 Databases : others

- 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 (DIP, ProNet, BIND, MINT), Protease db (MEROPS), biotechnology patents db, etc.;
- As well as many other resources concerning any aspects of macromolecules and molecular biology.



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; creation of db_new between releases (exemple: EMBLnew; TrEMBLnew....)
- Some servers add automatically useful cross-references to an entry (implicit links) in addition to already existing links (explicit links)



»»»»» Using sequences

Databases can be accessed easily using retrieval tools

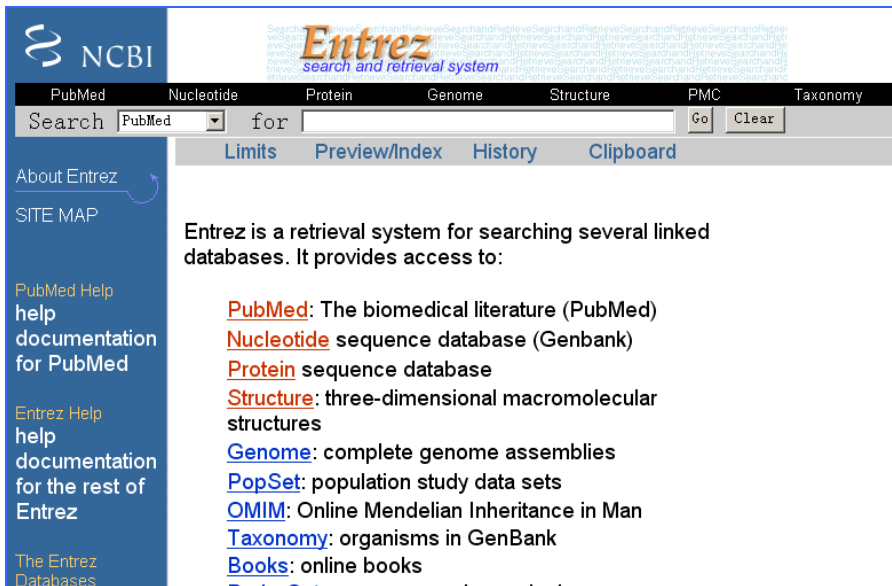
- **Sequence Retrieval System** (SRS, Europe) allows any flat-file db to be indexed to any other; allows to formulate queries across a wide range of different db types via a single interface, without any worry about data structure, query languages...
- **Entrez** (USA): less flexible than SRS but exploits the concept of « neighbouring », which allows related articles in different db to be linked together, whether or not they are cross-referenced directly
- **ATLAS**: specific for macromolecular sequences db (i.e. NRL-3D)
-



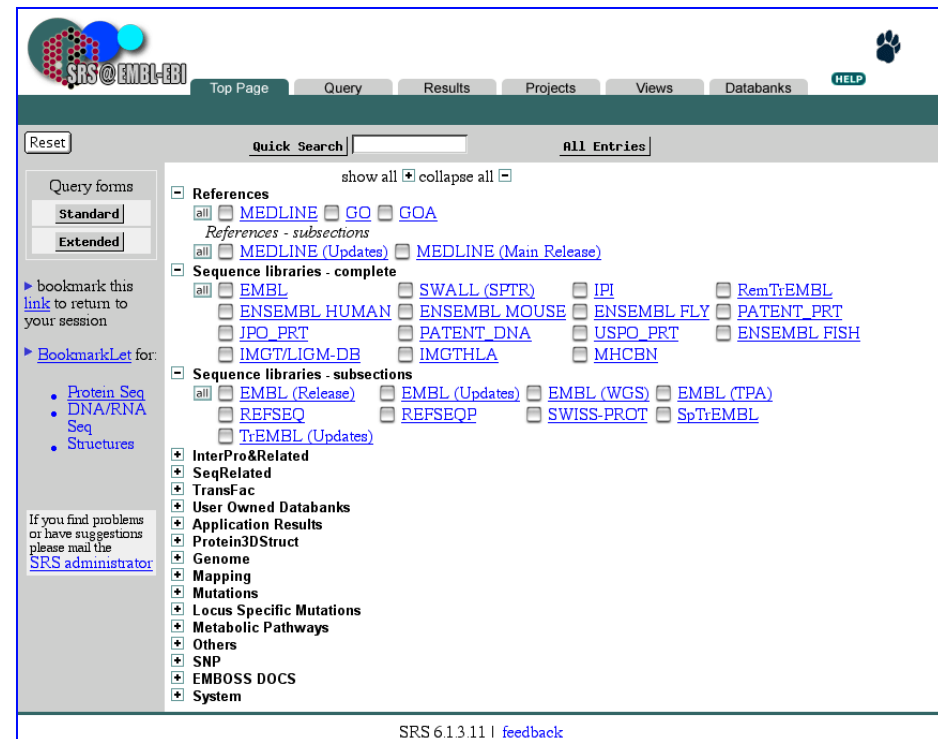
Information Retrieval System

Entrez (GenBank, NCBI)

SRS (Sequence retrieval System)(EMBL)



The screenshot shows the NCBI Entrez homepage. At the top, the NCBI logo is on the left, and the Entrez logo with the tagline "search and retrieval system" is on the right. Below the logo, there is a navigation bar with links to PubMed, Nucleotide, Protein, Genome, Structure, PMC, and Taxonomy. A search bar is present with a dropdown menu set to "PubMed" and a "Go" button. Below the search bar, there are links for "Limits", "Preview/Index", "History", and "Clipboard". On the left side, there is a sidebar with links for "About Entrez", "SITE MAP", "PubMed Help", "documentation for PubMed", "Entrez Help", "documentation for the rest of Entrez", and "The Entrez Databases". The main content area contains the text: "Entrez is a retrieval system for searching several linked databases. It provides access to:" followed by a list of databases: PubMed (The biomedical literature), Nucleotide (sequence database), Protein (sequence database), Structure (three-dimensional macromolecular structures), Genome (complete genome assemblies), PopSet (population study data sets), OMIM (Online Mendelian Inheritance in Man), Taxonomy (organisms in GenBank), and Books (online books).



The screenshot shows the SRS (Sequence Retrieval System) interface. At the top, there is a navigation bar with links to Top Page, Query, Results, Projects, Views, Databanks, and a HELP button. Below the navigation bar, there is a search bar with a "Reset" button and a "Quick Search" button. The main content area is divided into several sections: "Query forms" (Standard, Extended), "References" (all, MEDLINE, GO, GOA), "Sequence libraries - complete" (all, EMBL, ENSEMBL HUMAN, IPO_PRT, IMGT/LIGM-DB, SWALL (SPTR), ENSEMBL MOUSE, PATENT_DNA, IMGT/HLA, IPI, ENSEMBL FLY, USPO_PRT, MHCEN, RemTrEMBL, PATENT_PRT, ENSEMBL FISH), "Sequence libraries - subsections" (all, EMBL (Release), REFSEQ, TrEMBL (Updates), EMBL (Updates), REFSEQP, EMBL (WGS), SWISS-PROT, EMBL (TPA), SpTrEMBL), "InterPro&Related", "SeqRelated", "TransFac", "User Owned Databanks", "Application Results", "Protein3DStruct", "Genome", "Mapping", "Mutations", "Locus Specific Mutations", "Metabolic Pathways", "Others", "SNP", "EMBOSS DOCS", and "System". At the bottom, there is a footer with the text "SRS 6.1.3.11 | feedback".

