



**“I hope you’ve got a lot of disk space, Ted.
I think I accidentally just faxed you
the entire Internet.”**

An introduction to biological databases

Database or databank ?

- At the beginning, subtle distinctions were done between databases and databanks (in UK, but not in the USA), such as:
« Database management programs for the gestion of databanks »
- From now on, the term « database » (db) is usually preferred



What is a database ?

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



Databases: a « flat-file » example

« Introduction To Database » Teacher Database (ITDTdb) (flat file, 3 entries)

Accession number: 1

First Name: Amos

Last Name: Bairoch

Course: DEA=oct-nov-dec 2000

<http://expasy4.expasy.ch/people/amos.html>

//

Accession number: 2

First Name: Laurent

Last name: Falquet

Course: EMBnet=sept 2000;DEA=oct-nov-dec 2000;

//

Accession number 3:

First Name: Marie-Claude

Last name: Blatter Garin

Course: EMBnet=sept 2000;DEA=oct-nov-dec 2000;

<http://expasy4.expasy.ch/people/Marie-Claude.Blatter-Garin.html>

//

- Easy to manage: all the entries are visible at the same time !



Databases: a « relational » example

Relational database (« table file »):

Teacher	Accession number	Education
Amos	1	Biochemistry
Laurent	2	Biochemistry
M-Claude	3	Biochemistry



Course	Date	Involved teachers
DEA	Oct-nov-dec 2000	1,3
EMBnet	Sept 2000	2,3



Easier to manage; choice of the output



Why biological databases ?

- Explosive growth in biological data
- Data (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, as classical publications used to be !



Some statistics

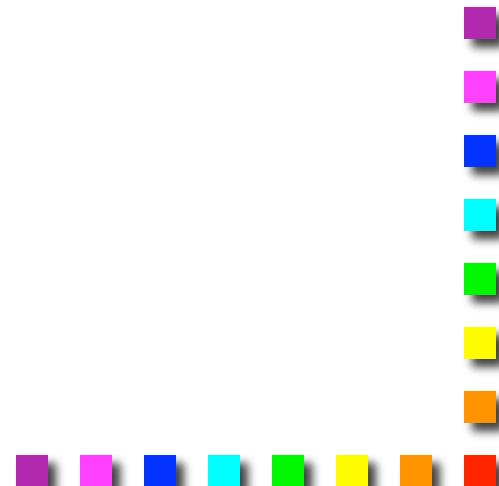
- More than 1000 different databases
- Variable size: <100Kb to >10Gb
 - DNA: > 10 Gb
 - Protein: 1 Gb
 - 3D structure: 5 Gb
 - Other: smaller
- Update frequency: daily to annually
- Generally accessible through the web (free!?)
 - Amos' links: www.expasy.org/alinks.html
 - Google: <http://www.google.com>



Biological databases

■ 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, DPInteract, ECDC,
ECGC, EC02DBASE, 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, HlVdb, 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) -> Primary db
- Genomics
- Protein domain/family → Secondary db
- Mutation/polymorphism
- Proteomics (2D gel, MS)
- 3D structure → Structure db
- Metabolism
- Bibliography
- Others (Microarrays)



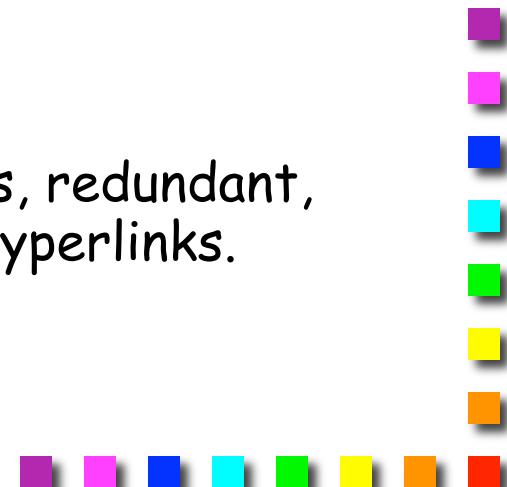
Distribution of sequence databases

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



Sequence Databases: some « technical » definitions

- Data storage management:
 - flat file: text file
 - relational (e.g., Oracle)
 - object oriented (rare in biological field)
- Format (flat file):
 - fasta
 - GCG
 - NBRF/PIR
 - MSF....
 - standardized format ?
- Federated databases: different autonomous, redundant, heterogeneous db linked together by links/hyperlinks.



Ideal minimal content of a « sequence » db

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



Sequence database: example

SWISS-PROT Flat file

taxonomy

ID EPO_HUMAN STANDARD; PRT; 193 AA.
AC P01588;
DT 21-JUL-1986 (Rel. 01, Created)
DT 21-JUL-1986 (Rel. 01, Last sequence update)
DT 30-MAY-2000 (Rel. 39, Last annotation update)
DE Erythropoietin precursor.
GN EPO.
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
RN [1]
RP SEQUENCE FROM N.A.
RX MEDLINE; 85137899.

reference

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).

...
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.

annotations

CC -!- SUBCELLULAR LOCATION: SECRETED.
CC -!- TISSUE SPECIFICITY: PRODUCED BY KIDNEY OR LIVER OF ADULT MAMMALS
AND BY LIVER OF FETAL OR NEONATAL MAMMALS.
CC -!- PHARMACEUTICAL: Available under the names Epogen (Amgen) and
CC Procrit (Ortho Biotech).
CC -!- DATABASE: NAME=R&D Systems' cytokine source book;
CC WWW="http://www.rndsystems.com/cyt_cat/epo.html".



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.

...
KW Erythrocyte maturation; Glycoprotein; Hormone; Signal; Pharmaceutical.
FT SIGNAL 1 27
FT CHAIN 28 193 ERYTHROPOIETIN.
FT PROPEP 190 193 MAY BE REMOVED IN PROCESSED PROTEIN.
FT DISULFID 34 188
...

...

Keywords



Sequence database: example (cont.)

sequence

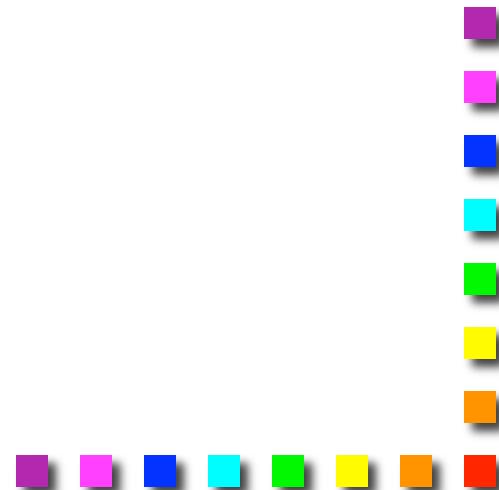
```
FT  DISULFID    34    188
FT  DISULFID    56     60
FT  CARBOHYD    51     51      N-LINKED (GLCNAC....).
FT  CARBOHYD    65     65      N-LINKED (GLCNAC....).
FT  CARBOHYD   110    110      N-LINKED (GLCNAC....).
FT  CARBOHYD   153    153
FT  CONFLICT     40     40      E -> Q (IN CAA26095).
FT  CONFLICT     85     85      Q -> QQ (IN REF. 5).
FT  CONFLICT   140    140      G -> R (IN CAA26095).
** Chromosomal location: 7q22
SQ  SEQUENCE 193 AA; 21306 MW; C91F0E4C26A52033 CRC64;
MGVHECPAWL WLLLSLLSLP LGLPVLGAPP RLICDSRVLE RYLLEAKEAE NITTGCAEH
SLNENITVPD TKVNFYAWKR MEVGQQAVEV WQGLALLSEA VLRGQALLVN SSQPWEPLQL
HVDKAVSGLR SLTLLRALG AQKEAISPPD AASAAPLRTI TADTFRKLFR VYSNFLRGKL
KLYTGEACRT GDR
//
```



Sequence database: example

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

```
>sp|P01588|EPO_HUMAN ERYTHROPOIETIN PRECURSOR - Homo sapiens (Human).  
MGVHECPAWLWLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAE  
NITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEA  
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTLLRALGAQKEAISPPD  
AASAAPLRTITADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR
```



Databases 1: nucleotide sequence

- 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



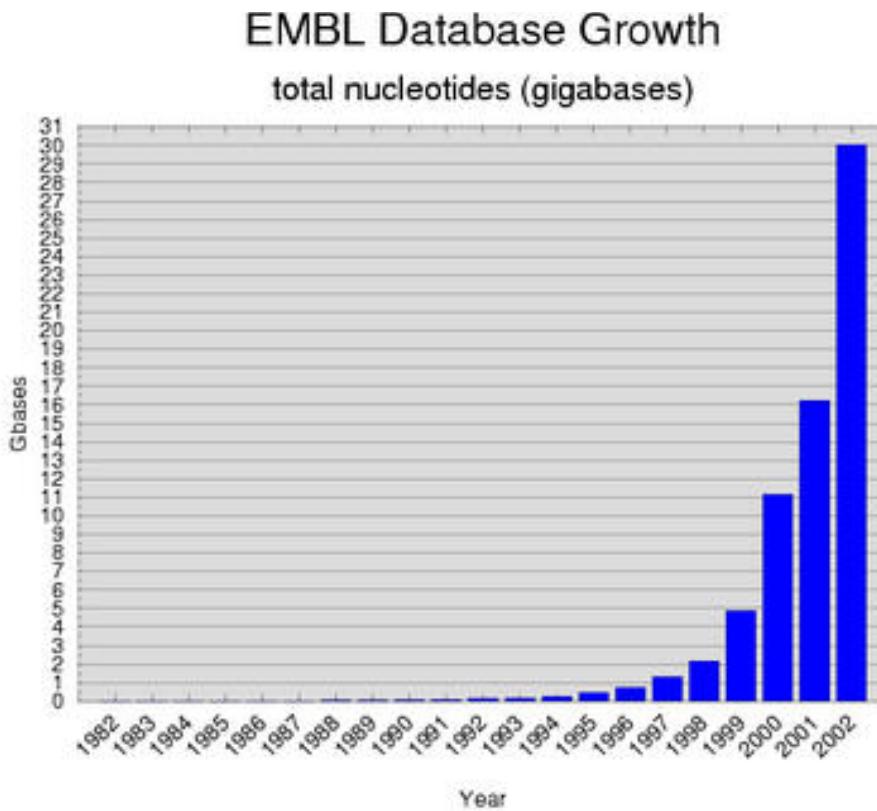
EMBL/GenBank/DDJB

- 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 and sequencing centers
 - Individual scientists
 - Patent offices (i.e. European Patent Office, EPO)
- Non-confidential data are exchanged daily
- Currently: 20×10^6 sequences, over 30×10^9 bp;
 - Stats: <http://www3.ebi.ac.uk/Services/DBStats/>
- Sequences from > 73' 000 different species;

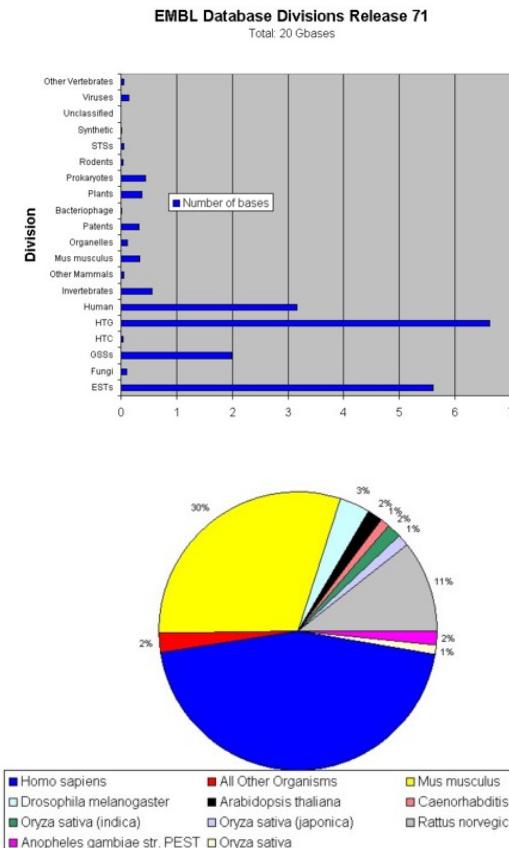


The tremendous increase in nucleotide sequences

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



1980: 80 genes fully sequenced !



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



EMBL/GenBank/DDJB

■ Unexpected informations you can find in these db:

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

■ Or:

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



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; taxonomy  
OC Eutheria; Primates; Catarrhini; Hominidae; Homo.  
XX  
RN [1]  
RP 1-3398  
RX MEDLINE; 85137899.  
RA Jacobs K., Shoemaker C., Rudersdorf R., Neill S.D., Kaufman R.J.,  
RA Mufson A., Seehra J., Jones S.S., Hewick R., Fritsch E.F., Kawakita M.,  
RA Shimizu T., Miyake T.;  
RT Isolation and characterization of genomic and cDNA clones of human  
RT erythropoietin;  
RL Nature 313:806-810(1985).  
XX  
DR GDB; 119110; EPO.  
DR GDB; 119615; TIMP1.  
DR SWISS-PROT; P01588; EPO_HUMAN.  
XX
```

keyword

taxonomy

references

Cross-references

...

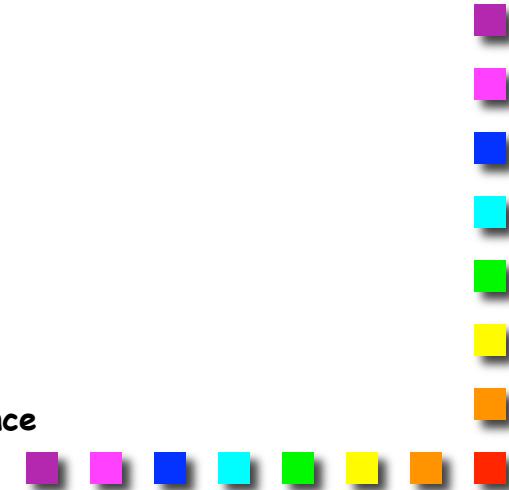


EMBL entry (cont.)

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=MGVHECPAWLWLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLLE
AKEAENITTCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRG
QALLVNSSQPEPLQLHVDKAVSGLRSLLRALGAQKEAISPPDAASAAPLRTITAD
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='untranslated region'
FT /number=5
XX
SQ Sequence 3398 BP; 698 A; 1034 C; 991 G; 675 T; 0 other;
agcttctggg cttccagacc cagctacttt gcggaaactca gcaacccagg catctctgag
tctccggcca agaccggat gccccccagg aggtgtccgg gagcccagcc tttcccatg 60
tctccggcca agaccggat gccccccagg aggtgtccgg gagcccagcc tttcccatg 120

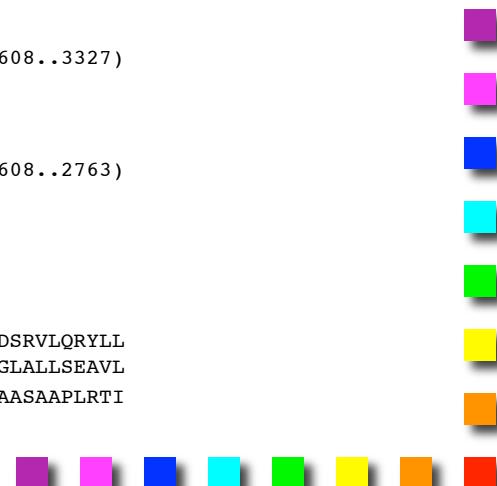
annotation

sequence



GenBank entry: example

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="MGVHECPAWLWLLSLLSPLGLPVLGAPPRLICDSRVLQRYILL
EAKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVL
RGQALLVNSSQPWEPLQLHVVDKAVSGLRSLLRALGAQEKAISPPDAASAAPLRTI
...



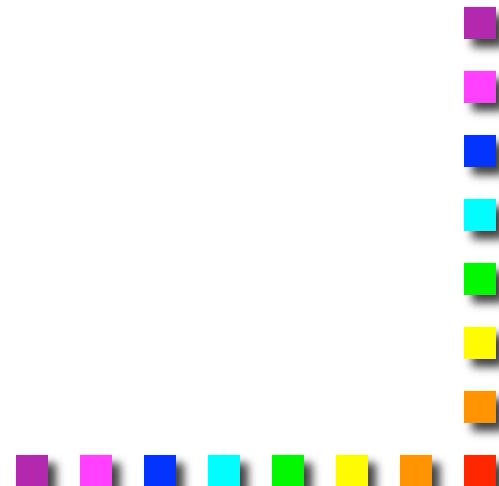
GenBank entry (cont.)

TADTFRKLFRVYSNFLRGKLKLYTGEACRTGDR"
inttron 628..1193
/number=1
exon 1194..1339
/number=2
mat_peptide join(1262..1339,1596..1682,2294..2473,2608..2760)
/product="erythropoietin"
inttron 1340..1595
/number=2
exon 1596..1682
/number=3
inttron 1683..2293
/number=3
exon 2294..2473
/number=4
inttron 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 agtttctggg cttccagacc cagctacttt gcggaactca gcaacccagg catctctgag
61 tctccggcca agaccggat gcccccagg aggtgtccgg gagcccgacc tttcccgat
121 agcagctccg ccagtcccaa gggtgcgcaa ccggctgcac tccccctcccg cgacccaggg
181 cccgggagca gcccccatga cccacacgca cgtctgcagc agccccgtca gccccggagc
241 ctcaacccag gcgtcctgcc cctgctctga cccccgggtgg cccctacccc tggcgacccc



DDJB entry: example

```
LOCUS      HSERPG      3398 bp    DNA          HUM      22-JUN-1993
DEFINITION Human gene for erythropoietin.
ACCESSION  X02158
VERSION    X02158.1
KEYWORDS   erythropoietin; glycoprotein hormone; hormone; signal peptide.
SOURCE     human.
ORGANISM   Homo sapiens
           Eukaryota; Metazoa; Chordata; Craniata; 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, 806-810(1985)
MEDLINE   85137899
COMMENT   Data kindly reviewed (24-FEB-1986) by K. Jacobs
FEATURES  Location/Qualifiers
source    1..3398
           /db_xref="taxon:9606"
           /organism="Homo sapiens"
mRNA      join(397..627,1194..1339,1596..1682,2294..2473,2608..3327)
CDS       join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
           /db_xref="SWISS-PROT:P01588"
           /product="erythropoietin"
           /protein_id="CAA26095.1"
           /translation="MGVHECPAWLWLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLLE
           AKEAENITTGCAEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRG
           QALLVNSSQPWEPLQLHVDKAVSGLRSLLRALGAQKEAISPPDAASAAPLRTITAD
           TFRKLFRVYSNFLRGKLKLYTGEACRTGDR »
...
...
```



DDJB (cont.)

```
mat_peptide      join(1262..1339,1596..1682,2294..2473,2608..2763)
                  /product="erythropoietin"
sig_peptide      join(615..627,1194..1261)
exon            397..627
                  /number=1
intron          628..1193
                  /number=1
exon            1194..1339
                  /number=2
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 agtttctggg cttccagacc cagctacttt gcggaactca gcaacccagg catctctgag
61 tctccgcccc agaccggat gccccccagg aggtgtccgg gagccccagcc tttcccagat
```



EMBL divisions

- EMBL has been divided into subdatabases to allow easier data management and searches
 - fun, hum, inv, mam, org, phg, pln, pro, rod, syn, unc, vrl, vrt
 - est, gss, htg, htc, sts, patent



EMBL: The Genome divisions

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



EMBL-EBI
European Bioinformatics Institute

EBI Home About EBI Research Services Toolbox Databases Downloads Submissions
GENOMES

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

GENOMES

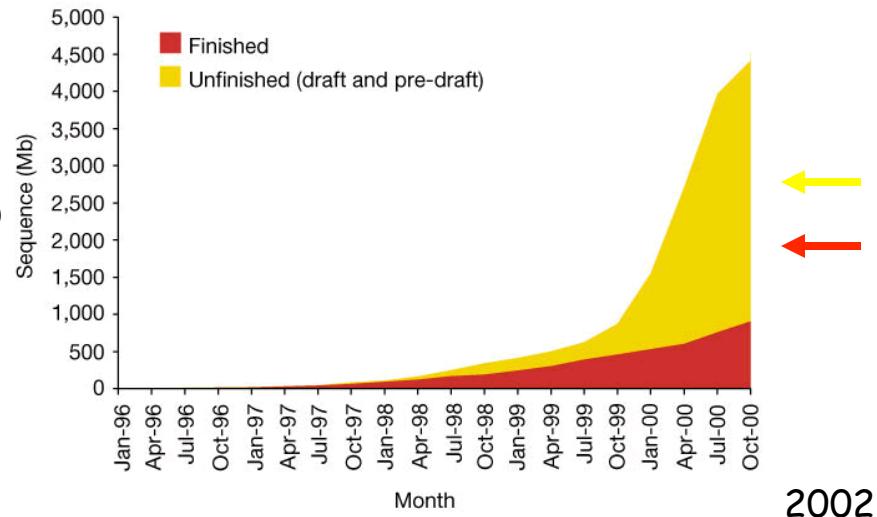

- Archaea
- Bacteria
- **Eukaryota** 
- Organelles
- Phages
- Plasmids
- Viroids
- Viruses

Other Links

- Proteomes
- Fasta3 Server
- Genomes MOT

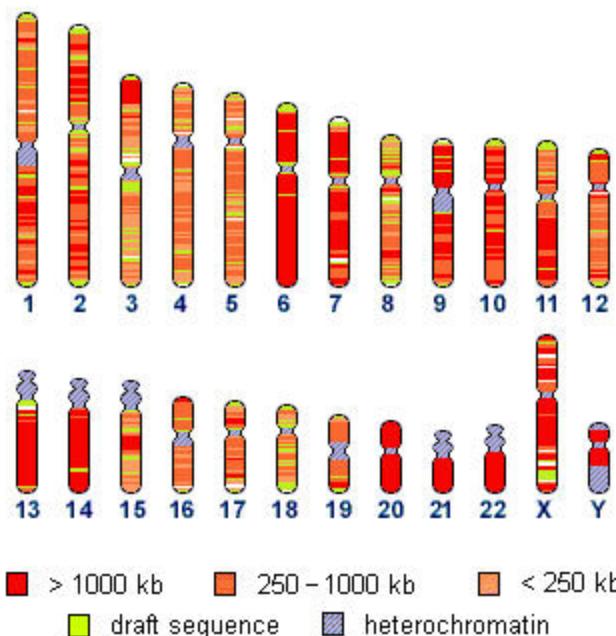
Human genome

- The completion of the draft human genome sequence has been announced on 26-June-2000.
- Publication of the public Human Genome Sequence in Nature the 15 th february 2001. Approx. 30,000 genes are analysed, 1.4 million SNPs and much more.
- The draft sequence data is available at EMBL/GENBANK/DDJB
- 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.
- The general assumption is that about 50% of the bases are redundant.



Human Genome Sequencing

Sequencing Progress



Genome Watch

31 Dec, 2001

Draft 34.8%

Finished 63.0%

Total 97.8%

Assumptions and additional statistics...

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.



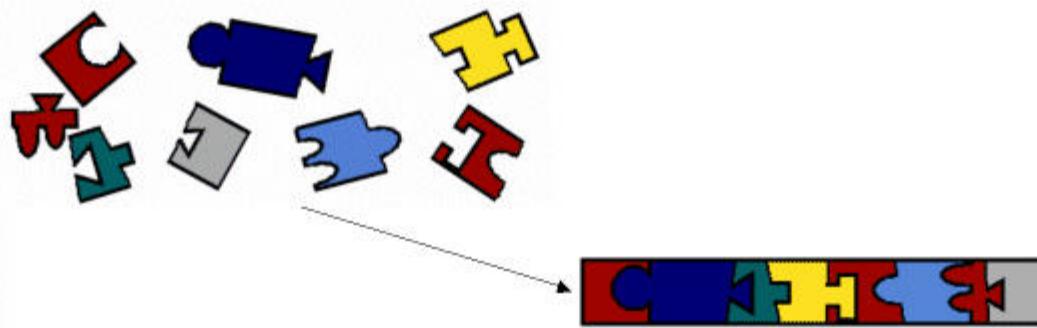
9 May, 2000

What's New

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

The Jigsaw Puzzle Genome

Modern DNA sequencing technology can only determine accurate sequences of short stretches of DNA (less than 1000 base pairs). Since the human genome is in excess of 3 billion base pairs long the genome has had to be sequenced in many small pieces that must be reassembled afterwards. The pieces are reassembled by comparing the sequence of the ends to find overlaps which can be used to join them together.



Nucleotid databases and « associated » genomic projects/databases

Problem:

Redundancy = makes Blasts searches of the complete databases useless for detecting anything beyond 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)...

PubMed Entrez BLAST OMIM Taxonomy Structure

Search LocusLink Display Brief Organism: All

Query: Go Clear

View Hs EPOR One of 1 Loci Save All Loci

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

[Click to Display mRNA-Genomic Alignments \(spanning 6544 bps\)](#)

PUB OMIM ACEVIEW UGENE MAP VAR HOMOL GDB

e! UCSC

Homo sapiens Official Gene Symbol and Name (HGNC)

EPOR: erythropoietin receptor

LocusID: 2057

Overview ?

RefSeq Summary: The erythropoietin receptor is a member of the cytokine receptor family. Upon erythropoietin binding, the erythropoietin receptor activates Jak2 tyrosine kinase which activates different intracellular pathways including Ras/MAP kinase, phosphatidylinositol 3-kinase and STAT transcription factors. The stimulated erythropoietin receptor appears to have a role in erythroid cell survival. Defects in the erythropoietin receptor may produce erythroleukemia and familial erythrocytosis.

Proteome Summary: Erythropoietin receptor, member of the cytokine receptor superfamily

Locus Type: gene with protein product, function known or inferred

Product: erythropoietin receptor precursor

Function [Submit GeneRIF](#) [\(All Pubs\)](#) ?

Phenotype: [Erythrocytosis, familial](#)

GeneRIF: Gene References into Function:

- [12118093](#) • functional significance of expression in breast cancer
- [12027890](#) • evidence for pY429pY431 being a new high affinity binding site for SOCS-3 on the EpoR
- [11884148](#) • Amino acid determinants of beta-hairpin conformation in erythropoietin receptor agonist peptides derived from a phage display library
- [12021194](#) • The extracellular binding site for ERP is now characterized. The site is located in the membrane proximal, extracellular part of the receptor. ERP binds to a region on the EPOR that contains the same sequence as ERP

Gene Ontology™:

Term	Evidence	Source	Pub
signal transduction	NR	Proteome	pm
erythropoietin receptor	E	Proteome	pm
integral plasma membrane protein	E	Proteome	pm

Other Ontologies:

Term	Evidence	Source	Pub
• Differentiation	NR	Proteome	pm
• Dendrite	E	Proteome	pm
• Other development	NR	Proteome	pm
• Cell fate specification	NR	Proteome	pm
• Integral membrane	NR	Proteome	pm
• Cell body (soma)	E	Proteome	pm
• Receptor (signalling)	NR	Proteome	pm

LocusLink / RefSeq Erythropoietin receptor

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
[LOCUSLINK](#)
Collaborators
Download
FAQ
[LOCUSLINK](#)

Collaborators
Download
FAQ
Help
Statistics
[RefSeq](#)

Relationships [?](#)

Mouse Homology Maps:

NCBI vs. MGD	9 5.00 cM	Epor	Hs Mm
UCSC vs. MGD	9 5.00 cM	Epor	Hs Mm
NCBI vs. EST-based RH Map	17 843.52 cR	AF237669	Hs Mm

Map Information [?](#)

Chromosome:	19	mv
Cytogenetic:	19p13.3-p13.2	RefSeq
Markers:	Chr. 19 D19S1015	D19S1015 mv
	Chr. 19 RH17696	mv
	Chr. - GDB:196998	
	Chr. - GDB:250953	
	Chr. - GDB:250955	
	Chr. - GDB:266625	

NCBI Reference Sequences (RefSeq) [?](#)

Category: REVIEWED

mRNA:	NM_000121
Protein:	NP_000112 erythropoietin receptor precursor BL
GenBank Source:	M60459,M76595

Category: NCBI Genome Annotation

Genomic	NT_011176	gb sv mv ev mm
Contig:		

Annotated transcripts/proteins for this locus:

Evidence:	supported by alignment with both mRNA and ESTs (3)
mRNA:	NM_000121
Protein:	NP_000112 BL

GenBank Sequences [?](#)

Nucleotide	Type	Protein	
M76595	g	AAA52393	BL
M77244	g	AAA52392	BL
S45332	g	AAB23271	BL
AK074082	m	BAB84908	BL
M34986	m	AAA52401	BL
M60459	m	AAA52403	BL
X57282	m	CAA40550	BL
X97671	m	CAA66260	BL

Additional Links [?](#)

- OMIM: [133171](#)
- UniGene: [Hs.127826](#)
- [GeneCards](#)

[To Top](#)

Questions or Comments?

Write to the [NCBI Service Desk](#)

[Disclaimer](#) [Privacy statement](#)

RefSeq a SWISS-PROT clone?

- The NCBI Reference Sequence project (RefSeq) will provide reference sequence standards for the naturally occurring molecules of the central dogma, from chromosomes to mRNAs to proteins. RefSeq standards provide a foundation for the functional annotation of the human genome. They provide a stable reference point for mutation analysis, gene expression studies, and polymorphism discovery.

- Molecule**

Complete Genome

- Accession Format**

NC_#####

- Genome**

Archaea, Bacterial,
Organelle, Virus, Viroid

Complete Chrom.

Eukaryote

Complete Sequence

Plasmid

Genomic Contig

Homo sapiens

mRNA

Homo sapiens, Mus musculus,
Rattus norvegicus



Protein

All of the above



mRNA

H. sapiens model transcripts



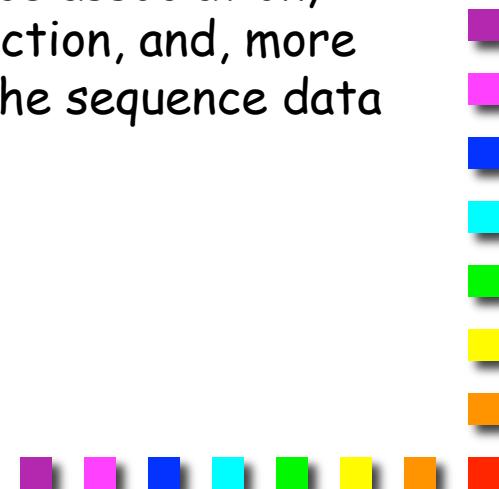
Protein

H. sapiens model proteins



RefSeq a SWISS-PROT clone?

- RefSeq records are created via a process consisting of:
 - identifying sequences that represent distinct genes
 - establishing the correct gene name-to-accession number association
 - identifying the full extent of available sequence data
 - creating a new RefSeq record with a status of:
 - PREDICTED (some part of the record is predicted)
 - PROVISIONAL (not yet reviewed by NCBI staff)
 - REVIEWED (reviewed and extended by NCBI staff)
 - Genome Annotation (contigs, mRNA and proteins generated automatically)
- Provisional RefSeq records are non-redundant and reviewed by a biologist who confirms the initial name-to-sequence association, adds information including a summary of gene function, and, more importantly, corrects, re-annotates, or extends the sequence data using data available in other GenBank records.



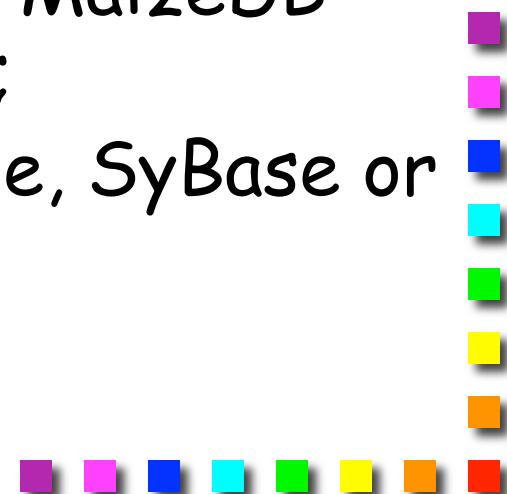
ESTs and Unigene

- Unigene is an ongoing effort at NCBI to cluster EST sequences with traditional gene sequences
- For each cluster, there is a lot of additional information included
- Unigene is regularly rebuilt. Therefore, cluster identifiers are **not** stable gene indices
- Species: Human, Mouse, Rat, Cow, Zebrafish, and recently also Frog, Cress, Rice, Barley, Maize, Wheat



Databases 2: genomics

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



MIM

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



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.

*133170 ERYTHROPOIETIN; EPO

Alternative titles; symbols

EP

TABLE OF CONTENTS

- [DESCRIPTION](#)
- [CLONING](#)
- [MAPPING](#)
- [ANIMAL MODEL](#)
- [REFERENCES](#)
- [SEE ALSO](#)
- [CONTRIBUTORS](#)
- [CREATION DATE](#)
- [EDIT HISTORY](#)

Database Links

[MEDLINE](#) [Protein](#) [DNA](#) [Genome](#) [LocusLink](#) [Gene Map](#) [GDB](#) [MDA](#) [Nomenclature](#)

Gene Map Locus: [7q21](#)

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

TEXT

DESCRIPTION

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

CLONING

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

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

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

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

MAPPING

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

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

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

ANIMAL MODEL

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

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

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

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

SEE ALSO

[Jacobs et al. \(1985\)](#) ; [Lin et al. \(1985\)](#) ; [Semenza et al. \(1989\)](#)

REFERENCES

Genecard

an electronic encyclopedia of biological and medical information
based on intelligent knowledge navigation technology

The GeneCards Encyclopedia integrates a subset of the information stored in major data sources dealing with **human genes and their products** (with a major focus on **medical aspects**). To facilitate access to this type of biomedical information, we continuously extract only those data that may be especially helpful for **efficient navigation** (e.g., have a look into the GeneCard for BRCA1, the famous breast cancer gene). The information you may presently find here includes:

- the official **name** approved by the UCL/HGNC Human Gene Nomenclature database
- a list of **synonyms** approved by HUGO, GDB, and/or SWISSPROT
- a list of the gene IDs in other gene-based resources (according to GDB, LocusLink, euGene, and/or HUGE)
- the (cytogenetic) **locus** of the gene (extracted from LocusLink, OMIM or UDB for all genes and additionally from CroW21 and Nature 405: 311-319 for the chromosome 21 genes) and its **genomic region** (visualized by the UCSC Golden Path, and/or Ensembl)
- the name of its **product(s)** (i.e., the **protein(s)** encoded according to the information stored in the gene); main features of this/these product(s), like cellular functions, expression patterns, similarities with other proteins, involvement in diseases (extracted from SWISS-PROT, MIPS, and/or BLOCKS)
- The UniGene cluster of **sequences** related to the gene, stored in GenBank
- homologous genes in the **mouse** and **worm** (orthologs; extracted from MGD and Stony Brook C.elegans-H.sapiens Alignment Database)
- a list of **disorders** and **mutations** in which the gene is involved according to genetic evidence (extracted from OMIM, SWISS-PROT, HGMD (Human Gene Mutation Database), BCGD (Breast Cancer Gene Database), TGDB (Tumor Gene Database), and/or GeneClinics)
- The **coordinates** as distance from the p terminus of the chromosome (in megabases) as calculated by UDB
- Titles of related **research articles** with links to the abstract and full citation in PubMed
- **Medical applications**, like new therapies and diagnoses, that are based on knowledge about this gene (extracted from The Doctor's Guide)

Links to all the sources where the information was extracted from are always given (we use links to the respective entries in those databases, in addition to their homepages, so that users are taken **directly** to the **detailed information**).

Release 1.0 beta
20 Dec 2001
32057 records

A portal to the human genome

[Text search](#)[BLAST search](#)[GeneLynx guide](#)[GeneLynx info](#)

View GeneLynx record

Enter a GeneLynx ID.

ID: Go

[GeneLynx Home](#)[Text Search](#)[BLAST search](#)[Linking to GeneLynx](#)[Resource submission](#)[GeneLynx guide](#)[GeneLynx info](#)

GeneLynx is a portal to a collection of hyperlinks for each human gene. It is implemented as an easily extensible relational database with a straightforward user interface.

You can access the information about a particular human gene by providing any reasonable identifier - just type a keyword, ANY accession number or ID below, or submit a related protein or nucleotide sequence on the BLAST search page. You can also perform a more refined keyword search on the Text search page.

**Parts of GeneLynx were out of function January 11-13, 2002 due to server misconfiguration.
We apologize for the inconvenience.**

Quick Search

Enter one or more terms separated by spaces.

 Go Clear

Combine terms with: AND OR

Exclude low-scoring hits

Send comments and questions to [Boris Lenhard](#)

Release 1.1
07 May 2002
32226 records

View GeneLynx
record
Enter a GeneLynx ID,
ID: Go

GeneLynx Home

Text Search

BLAST search

Linking to
GeneLynx

Resource
submission

GeneLynx guide

GeneLynx info

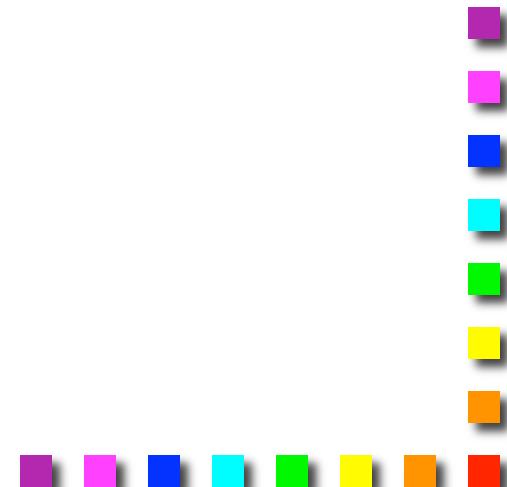
GeneLynx #5230

Gene name	EPO
Description	erythropoietin
Locus	7q22
Submit comment for this GeneLynx record	
<hr/>	
<i>Summary pages</i>	
LocusLink	2056
GeneCards	EPO
Unigene	Hs.2303
Swiss_Prot	EPO_HUMAN
KEGG gene	2056
EGAD	3760
euGenes	HUGN0002056
MIPS	771
HumanPSD	EPO
<hr/>	
<i>Genomic resources</i>	
Genomic sequences	NCBI EBI DBG AF053356 NCBI EBI DBG G20209 NCBI EBI DBG M11319
	NCBI EBI DBG X02158
GDB	119110
GenAtlas	EPO
Ensembl gene	ENSG00000087083 ENSG00000130427

Collections of hyperlinks for each human gene

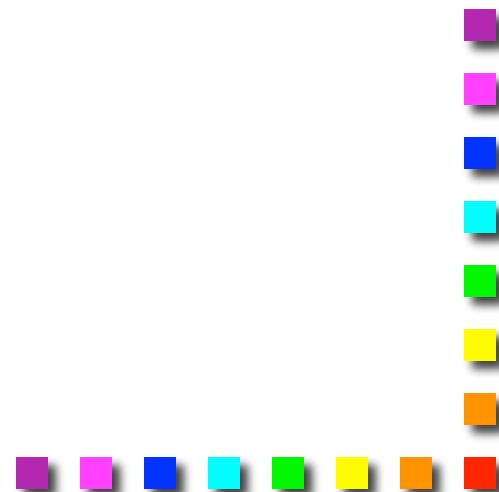
Ensembl

- Contains all the human genome DNA sequences currently available in the public domain.
- Automated annotation: by using different software tools, features are identified in the DNA sequences:
 - Genes (known or predicted)
 - Single nucleotide polymorphisms (SNPs)
 - Repeats
 - Homologies
- Created and maintained by the EBI and the Sanger Center (UK)
- www.ensembl.org



Databases 3: mutation/polymorphism

- Contain informations on sequence variations that are linked or not to genetic diseases;
- Mainly human but: OMIA - Online Mendelian Inheritance in Animals
- **General db:**
 - OMIM
 - HGMD - 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
 -



Mutation/polymorphisms: definitions

- SNPs: single nucleotide polymorphisms
 - 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 mutation depends on the db (aa no 1 is not necessary the initiator Met !)



Mutation/polymorphisms

■ dbSNP consortium <http://snp.cshl.org/>

- Bayer, Roche, IBM, Pfizer, Novartis, Motorola.....
- Mission: develop up to 300,000 SNPs distributed evenly throughout the human genome and make the informations related to these SNPs available to the public without intellectual property restrictions. The project started in April 1999 and is anticipated to continue until the end of 2001.

■ 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
- Aug 24, 2000 , dbSNP has submissions for 803557 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



Mutation/polymorphisms

- Generally modest size; lack of coordination and standards in these databases making it difficult to access the data.
- There are initiatives to unify these databases
Mutation Database Initiative (4th July 1996).
- SVD - Sequence Variation Database project at EBI (HMutDB)
 - <http://www2.ebi.ac.uk/mutations/>
- HUGO Mutation Database Initiative (MDI).
Human Genome Variation Society
 - <http://www.genomic.unimelb.edu.au mdi/dblist/dblist.html>



Human Genome Variation society



Variation Databases and Related Sites

[Join HGVS](#)[FAQ](#)[Variation Databases and Related Sites](#)[Recommendations](#)[Waystation/Central Database project](#)[Meeting Reports](#)[Newsletters](#)[Meetings of Interest](#)[Relevant Publications](#)

This list was compiled by Rania Horaitis at the [Genomic Disorders Research Centre](#), Melbourne. Any errors, changes or additions, including "dead" sites should be reported to horaitis@medstv.unimelb.edu.au

[Disclaimer](#)[Locus Specific Mutation Databases](#)[Disease Centered Central Mutation Databases](#)[Central Mutation & SNP Databases](#)[National & Ethnic Mutation Databases](#)[Clinical & Patient Aspects Databases](#)[Non Human Mutation Databases](#)[Artificial Mutations Only](#)[Other Related Databases](#)[Education Resources for Teachers & Students](#)

GENE DESIGNATION/ OMIM No.	DATABASE NAME	INTERNET ADDRESS	CURATOR/S
ABCD1 300100	Mutation Database for X-linked Adrenoleukodystrophy ENCOURAGED BY HUGO/MDI	http://www.x-ald.nl	Hugo W. Moser Dept. of Neurogenetics Kennedy Krieger Inst. Baltimore, USA Stephan Kemp Ronald R.J.A. Wanders Lab. of Genetic Metabolic Diseases Academic Medical Ctr Amsterdam, The Netherlands.
ABCR 601691	Keio Mutation db for Eye Disease Genes (KMEYEDB)	http://mutview.dmb.med.keio.ac.jp	S. Minoshima S. Mitsuyama S. Ohno T. Kawamura N. Shimizu Keio Univ. School of Med. Tokyo, Japan
ABO 110300 ACHE 100740	Blood Group Antigen Mutation Database ENCOURAGED BY HUGO/MDI	http://www.bioc.aecom.yu.edu/bgmut/index.htm	Olga O. Blumenfeld Department of Biochemistry, Santosh Patnaik, Department of Cell Biology, Albert Einstein College of Medicine New York, NY, U.S.A
ADL 600119	alpha-Sarcoglycan	http://www.dmd.nl/sgca_home.html	Johan T. den Dunnen Bert E. Bakker Leiden Univ. Med. Centre Leiden, The Netherlands
ADSL 103050	Adenylosuccinate Lyase Mutations Database Home Page	http://www.icp.ucl.ac.be/adslbd/	Johan T. den Dunnen Bert E. Bakker Leiden Univ. Med. Centre Leiden, The Netherlands
ADRB3 109691	Obesity at GeneDis (beta-3-adrenergic receptor)	http://life2.tau.ac.il/GeneDis/Tables/Obesity/obesity.html	Rachel Kreisberg-Zakarin, Bioinformatics Unit, Tel-Aviv University, Israel
ALB 103600	Albumin mutation database	http://www.albumin.org/	Eugene W. Holowachuk The Mary Imogene Bassett Hospital Research Institute Cooperstown, NY, U.S.A.



This site is sponsored in part by Serologicals.

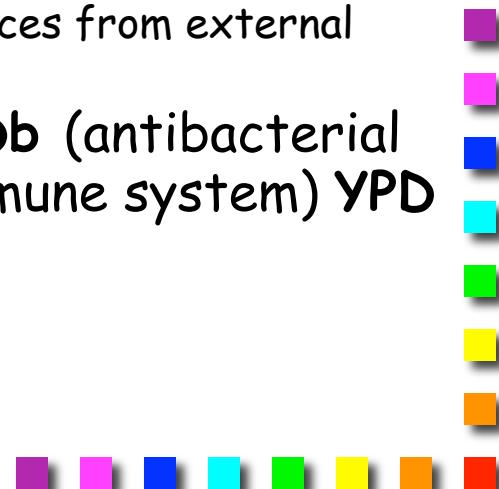


Human Serum Albumin Mutations

Residue	Amino acid chg.		Codon change (lower case) (+ = known)	Geographical names and [references]; in order of reports
	From	To		
-2	Arg	His	CgT -> CaT +	Lille [1]; Pollibauer, Somalia, Tokushima [2], Taipei [3], Fukuoka-2 [4], Varese [5]; Wu Yang [6]; Mayo, Komagone-3* [7].
-2@	Arg	Cys	cGT -> tGT +	(-2 to -6 omitted) Malmo I [8], Kaikoura [9], Tradate [5] Redhill (see also residue 320)[10,11]; high in Italy, Sweden, 3 homozygotes*[12]; Ilidut [62].
-1	Arg	Gln	CgA -> CaA +	Christchurch [13]; Gainesville [14, 3], Y, Honolulu-2[4], Fukuoka-3 [4], Mayo* [7]; Shizuoka [15]
-1	Arg	Pro	CgA -> CcA	Takefu [3]; Honolulu-1 [3]
-1	Arg	Leu	CgA -> CtA	Jaffna [16]

Database 4: protein sequence

- SWISS-PROT: created in 1986 (A.Bairoch)
- TrEMBL: created in 1996; complement to SWISS-PROT; derived from automated EMBL CDS translations (« proteomic » version of EMBL)
- PIR-PSD: Protein Information Resources
<http://pir.georgetown.edu/>
- All together a new unified database: UniProt??
- GenPept: derived from automated GenBank CDS translations and journal scans (« proteomic » version of GenBank)
- MIPS: Martinsried Institute for Protein Sequences
 - PIR + PATCHX (supplement of unverified protein sequences from external sources)
- Examples: **NRL-3D** from PDB (3D struture), **AMSDb** (antibacterial peptides), **GPCRDB** (7 TM receptors), **IMGT** (immune system) **YPD** (Yeast) etc.



SWISS-PROT

- Collaboration between the SIB (CH) and EMBL/EBI (UK)
- 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



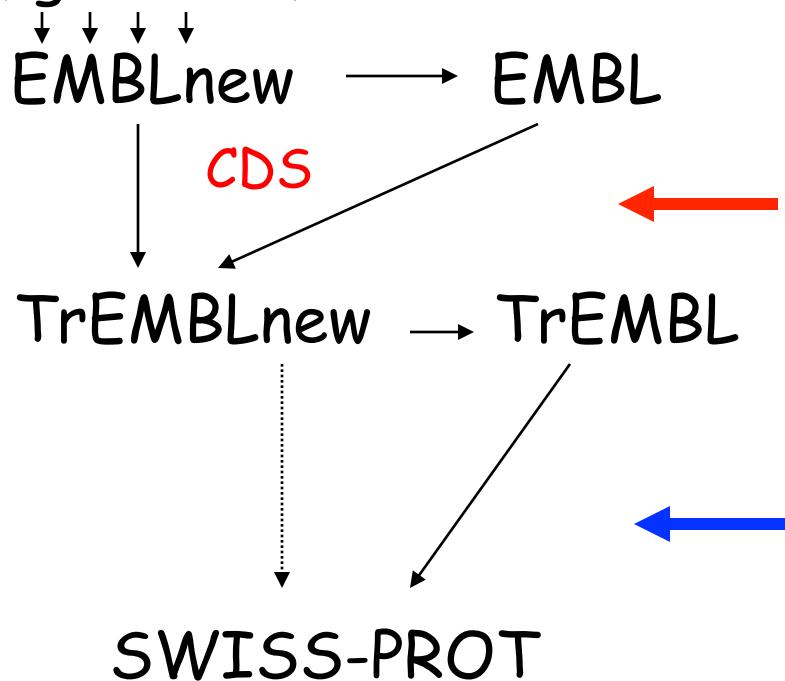
TrEMBL (Translation of EMBL)

- Computer-annotated supplement to SWISS-PROT, as it is impossible to cope with the flow of data...
- Well-structured SWISS-PROT-like resource
- Derived from automated EMBL CDS translation (maintained at the EBI (UK))
- TrEMBL is automatically generated and annotated using software tools (incompatible with the SWISS-PROT in terms of quality)
- TrEMBL contains all what is **not yet** in SWISS-PROT
- Yerk!! But there is no choice and these software tools are becoming quite good !



The simplified story of a Sprot entry

cDNAs, genomes,



« Automatic »

- Redundancy check (merge)
- InterPro (family attribution)
- Annotation

« Manual »

- Redundancy (merge, conflicts)
- Annotation
- Sprot tools (macros...)
- Sprot documentation
- Medline
- Databases (MIM, MGD....)
- Brain storming

Once in Sprot, the entry is no more in TrEMBL, but still in EMBL (archive)



SWISS-PROT introduces a new arithmetical concept !

How many sequences in SWISS-PROT + TrEMBL ?

113'000 + 670'000 ≈ about 450'000

(sept 2002)

- SWISS-PROT and TrEMBL (SPTR)
a minimal of redundancy



TrEMBL divisions

- TrEMBL: SPTrEMBL + REMTrEMBL
 - SPTrEMBL: TrEMBL entries that will eventually be integrated into SWISS-PROT, but that have not yet been manually annotated
 - REMTrEMBL: sequences that are not destined to be included in SWISS-PROT
 - Immunoglobulins and T-cell receptors
 - Synthetic sequences
 - Patented sequences
 - Small fragments (<8 aa)
 - CDS not coding for real proteins
 - TrEMBL new: updates to the latest release of TREMBL
- SPTR (SWall) = SWISS-PROT + (SP)TrEMBL + TrEMBLnew



TrEMBL divisions

■ Subdivisions

■ Archae	arc
■ Fungus	fun
■ Human	hum
■ Invertebrate	inv
■ Mammals	mam
■ Major Hist. Comp.	mhc
■ Organelles	org
■ Phage	phg
■ Plant	pln
■ Prokaryote	pro
■ Rodent	rod
■ Uncommented	unc
■ Viral	vrl
■ Vertebrate	vrt



	Line code	Content	Occurrence in an entry
taxonomy	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
	OS	Organism species	One or more
	OG	Organelle	Optional
references	OC	Organism classification	One or more
	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 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

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

Entry name
Accession
number

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC 075144; Q9NRQ4; 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 G150)
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., Zolnier 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-Joannopolulos K.;
RT "G150 molecules and uses thereof.";
RL Patent number WO0121796, 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; AAC34739.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF289028; AAC01176.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; ABO14553; BAA31628.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
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; SM004009; 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 GHW -> FSHMLLILS (IN ISOFORM 2)
SQ SEQUENCE 302 AA; 3349 NM; 647934E21B55E34A CRC64:
MRLGSPGLF LLFSSLRADT QEKEVRAMVG SDVELSCAP EGSRFDLNDV VVYVQTSSESK
TVTYTHIPQN SSLENVDSRY RNRLAHMSPAQ MLRGDFSLRL FNVTQPDQEKG FHCLVLSQL
GFEQVLSEPV TLHVAANFSV PVWSAPHPS QDELTTCTS INGYVPRNVY WINKNTDSLL
DQLQNDTFV LMHGRGLYDVY SVLRVIARTPS VNIGCCEIVN LLQQMLTVGS QTGDNGIERD
KITEPNPVSTG ERKNAATWSIL AVLCLVVVA VAIIGHVCRD CLQHSYAGAW AVSPTELTG
HV

sequence

SWISS-PROT: O75144

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

Protein name
Gene name
Taxonomy

Red arrows point from the taxonomy section to the protein name, gene name, and taxonomy sections of the SWISS-PROT entry.

ID ICOL_HUMAN **STANDARD**; **PRT**; 302 AA.
AC O75144; Q9NRQ1; QSHD18;
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 G150)
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 Tax ID=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 Yoshihaga 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., Zolnier 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-Joannopolulos K.;
RT "G150 molecules and uses thereof.";
RL Patent number WO0121796, 29-MAR-2001.

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

CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch).
CC -----
DR EMBL; AF199028; AAC34739.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; ABO14553; BAA31628.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
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 -> ESWNLLLLS (IN ISOFORM 2).
SQ SEQUENCE 302 AA; 33349 MW; 647934E21B55E34A CRC64;
MRLGSPGLF LLFSSLRADT QKEVFRAMVG SDVELSCAP EGSRFDLNDV VVWQTSSESK
TVTYTHIPQN SSLENVDSRY RNRLHMLSPAG MLRGDFSLRL FNVTQPCDEQK FHCLVLSQL
GFEQVLSEVEV TLHVAANFSV PVWSAPHPS QDELTTCTCS INGYPRPNVY WINKNTNSLL
DQLQNDTFV LMNRGLYDVV SVLRNIARTPS VNIGCCEVN LQQMLNTVGs QTGDNGIERD
KITEPNPVSTG EKNAATWSIL AVLCLVVVVA VAIGHVCRDR CLQHSYAGAW AVSPTELTG
//

SWISS-PROT: O75144

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

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; QSHD18;
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 G150) (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 Yoshihaga 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., Zolnier 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-Joannopolulos K.;
RT "G150 molecules and uses therefor.";
RL Patent number WO0121796, 29-MAR-2001.

CC --!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC --!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC --!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC --!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART, KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES, SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC --!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC --!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG 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 onward for an unknown reason.
CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch).
CC -----
DR EMBL; AF199028; AAC34739.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF289028; AAC01176.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; ABO14553; AAC31628.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
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 HUGO; 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 -> ESWNLLLLS (IN ISOFORM 2).
SQ SEQUENCE 302 AA; 3349 MW; 647934E21B55E34A CRC64;
MRLGSPGLLF LLFSSLRADT QKEVFRAMVG SDVELSCAP EGSRFDLNDV VVWQTSSESK
TVTYTHIPQN SSLENVDSRY RNRLHMLSPAG MLRGDFSLRL FNVTQPDQEKF FHCLVLSQL
GFEQVLSEVEV TLHVAANFSV PVWSAPHPS QDELFTCTS INGYVRPNVY WINKNTDSLL
DQLQNDTFV LMHRLGRYDVY SVLRRIARTPS VNIGCCEVN LLQQNLTVG5 QTGDNGIERD
KITEENPVSTG EKNAATWSIL AVLCLVVVVA VAIGHVCRDR CLQHSYAGAW AVSPTELTG
HV //

References

Comments

SWISS-PROT: O75144

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

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; QSHD18;
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 G150) (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; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
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 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 Yoshihaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K., Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H., 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 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., Leonard J.P., Hunter S.E., Zolnier R., Thomas J.L., Miyashiro J.S., Jacobs K.A., Collins M.;
RT "Identification of GL50, a novel B7-like protein that functionally 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., Kotani H., Nomura N., Ohara O.;
RT "Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can 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-Joannopolulos K.;
RT "G150 molecules and uses thereof.";
RL Patent number WO0121796, 29-MAR-2001.

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

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

DR EMBL; AF199028; AAC034739.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF289028; AAC01176.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF216749; AAC16241.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; ABO14553; BAA3628.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
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 -> ESWNLLLLS (IN ISOFORM 2).
SQ SEQUENCE 302 AA; 33349 MW; 647934E21B55E34A CRC64;
MRLGSPGLF LLFSSLRADT QKEVFRAMVG SDVELSCACP EGSRFDLNDV VVWQTSSESK
TVTYTHIPQN SSLENVDSRY RNRLHMLSPAG MLRGDFSLRL FNVTQPCDEQK FHCLVLSQL
GFEQVLSEPV TLHVAANFSV PVWSAPHPS QDELTTCTS INGYVRPNVY WINKNTDSLL
DQLQNDTFV LMHGRGLYDV SVLRNIARTPS VNIGCCEVN LLQQLNTVGs QTGDNGIERD
KITEENPVSTG EKNAATWSIL AVLCLVVVA VAIIGHVCRD CLQHSYAGAW AVSPTELTG
HV //

SWISS-PROT: O75144

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

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; QSHD18;
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 G150) (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; 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] Wang S., Zhu G., Chapoval A.I., Dong H., Tamada K., Ni J., Chen L.; "Costimulation of T cells by B7-H2, a B7-like molecule that binds 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] Yoshihaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K., Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H., Hui A., Kohno T., Manoukian R., Whoriskey J.S., Coccia M.A.; "Characterization of a new human B7-related protein: B7RP-1 is the 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] Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A., Leonard J.P., Hunter S.E., Zolnier R., Thomas J.L., Miyashiro J.S., Jacobs K.A., Collins M.; "Identification of GL50, a novel B7-like protein that functionally 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] Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A., Kotani H., Nomura N., Ohara O.; "Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can 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.; "G150 molecules and uses thereof.";
RL Patent number WO0121796, 29-MAR-2001.

Cross-references

CC --!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC --!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC --!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC --!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART, KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES, SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC --!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC --!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG 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 onward for an unknown reason.
CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch).
CC -----
DR EMBL: AF199028; AAC34739.1; -. [EMBL / GenBank / DDBJ / CoCodingSequence]
DR EMBL: AF289028; AAG01176.1; -. [EMBL / GenBank / DDBJ / CoCodingSequence]
DR EMBL: AF216749; AAK16241.1; -. [EMBL / GenBank / DDBJ / CoCodingSequence]
DR EMBL: ABO14553; BAA31628.1; ALT_SEQ. [EMBL / GenBank / DDBJ / CoCodingSequence]
DR EMBL: AX100595; CAC36465.1; -. [EMBL / GenBank / DDBJ / CoCodingSequence]
DR MIM: 605717; -. [NCBI / EBI]
DR GeneCards: ICOSL
DR GeneLynx: ICOSL
DR Ensembl: O75144.
DR InterPro: IPRO03599; Ig.
DR InterPro: IPRO03006; Ig MHC.
DR InterPro: IPRO03600; 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: GFT 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 -> ESWNLLLLS (IN ISOFORM 2).
SQ SEQUENCE 302 AA; 33349 MW; 647934E21B55E34A CRC64;
MRLGSPGLF LLFSSLRADT QKEVFRAMVG SDVELSCACP EGSRFDLNDV VVWQTSSESK TVTUVTYHQPON SSLENVDSRY RNRLHMLSPAG MLRGDFSLRL FNVTQPCDEQK FHCLVLSQL GFOQEVLSSVEV TLHVAANFSV PVVSAPHSFS QDELFTTCFSV INGYVRPNVY WINKTDSNL DQLQNDTFV LMHGRGLYDVY SVLRILARTPS VNIGCCEVN LLQQNLTVG5 QTGNDIGERD KITENPVSTG EKNAATWSIL AVLCLVVVVA VAIGHVCRDR CLQHSYAGAW AVSPTELTG HV

SWISS-PROT: O75144

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

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; QSHD18;
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 G150) (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; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
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 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 Yoshihaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K., Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H., 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 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., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A., Leonard J.P., Hunter S.E., Zolnier R., Thomas J.L., Miyashiro J.S., Jacobs K.A., Collins M.;
RT "Identification of GL50, a novel B7-like protein that functionally 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., Kotani H., Nomura N., Ohara O.;
RT "Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can 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 "G150 molecules and uses thereof.";
RL Patent number WO0121796, 29-MAR-2001.

Keywords

CC --!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC --!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC --!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC --!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART, KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES, SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC --!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC --!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG 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 onward for an unknown reason.
CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch).
CC -----
DR EMBL: AF199028; AAC34739.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL: AF289028; AAG01176.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL: AF216749; AAK16241.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL: ABO14553; BAA31628.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL: AX100595; CAC36465.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR MIM: 605717; -. [NCBI / EBI]
DR GeneCards: ICOSL
DR GeneLynx: ICOSL
DR Ensembl: O75144.
DR InterPro: IPRO03599; Ig.
DR InterPro: IPRO03006; Ig MHC.
DR InterPro: IPRO03600; 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: GFT 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 -> ESWNLLLLS (IN ISOFORM 2).
SQ SEQUENCE 302 AA; 33349 MW; 647934E21B55E34A CRC64;
MRLGSPGLF LLFSSLRADT QKVEKVRAMVG SDVELSCACP EGSRFDLNDV VVWVQTSSESK
TVVTHYIPQN SSLENVDSRY RNRLALMSPAG MLRGDFSLRL FNVTQPDQEKF FHCLVLSQL
GFEQVLSEPV TLHVAANFSV PVWSAPHSF QDELTTCTCS INGYVPRNVY WINKTDSNL
DQLQNDTFV LMHGRGLYDV SVLRILARTPS VNIGCCEVN LLQQMLTVGS QTGNDIGERD
KITEPNVSTG EKNAATWSIL AVLCLVVVVA VAIGHVCRDR CLQHSGAYAG AVSPTELTG
HV //

SWISS-PROT: O75144

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

ID ICOL_HUMAN STANDARD; PRT; 302 AA.
AC O75144; Q9NRQ1; QSHD18;
DT 15-OCT-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 G150) (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; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
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] 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 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] Yoshihaga S.K., Zhang M., Pistillo J., Horan T., Khare S.D., Miner K., Sonnenberg M., Boone T., Brankow D., Dai T., Delaney J., Han H., 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 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] Ling V., Wu P.W., Finnerty H.F., Bean K.M., Spaulding V., Fouser L.A., Leonard J.P., Hunter S.E., Zolnier R., Thomas J.L., Miyashiro J.S., Jacobs K.A., Collins M.; RT "Identification of GL50, a novel B7-like protein that functionally 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] Ishikawa K.-I., Nagase T., Suyama M., Miyajima N., Tanaka A., Kotani H., Nomura N., Ohara O.; RT "Prediction of the coding sequences of unidentified human genes. X. The complete sequences of 100 new cDNA clones from brain which can 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 "G150 molecules and uses thereof.";
RL Patent number WO0121796, 29-MAR-2001.

Feature table (sequence description)

CC --!- FUNCTION: LIGAND FOR THE T-CELL-SPECIFIC CELL SURFACE RECEPTOR ICOS. ACTS AS A COSTIMULATORY SIGNAL FOR T-CELL PROLIFERATION AND CYTOKINE SECRETION; INDUCES ALSO B-CELL PROLIFERATION AND DIFFERENTIATION INTO PLASMA CELLS. COULD PLAY AN IMPORTANT ROLE IN MEDIATING LOCAL TISSUE RESPONSES TO INFLAMMATORY CONDITIONS, AS WELL AS IN MODULATING THE SECONDARY IMMUNE RESPONSE BY CO-STIMULATING MEMORY T-CELL FUNCTION (BY SIMILARITY).
CC --!- SUBCELLULAR LOCATION: Type I membrane protein (By similarity).
CC --!- ALTERNATIVE PRODUCTS: AT LEAST 2 ISOFORMS; 1 (SHOWN HERE) AND 2; ARE PRODUCED BY ALTERNATIVE SPLICING.
CC --!- TISSUE SPECIFICITY: ISOFORM 1 IS WIDELY EXPRESSED (BRAIN, HEART, KIDNEY, LIVER, LUNG, PANCREAS, PLACENTA, SKELETAL MUSCLE, BONE MARROW, COLON, OVARY, PROSTATE, TESTIS, LYMPH NODES, LEUKOCYTES, SPLEEN, THYMUS AND TONSIL), WHILE ISOFORM 2 IS DETECTED ONLY IN LYMPH NODES, LEUKOCYTES AND SPLEEN.
CC --!- INDUCTION: CONSTITUTIVE EXPRESSION IS FURTHER ENHANCED BY TREATMENT WITH TNF-ALPHA IN PERIPHERAL BLOOD B-CELLS AND MONOCYTES, WHILE IT IS DECREASED IN DENDRITIC CELLS.
CC --!- SIMILARITY: BELONGS TO THE IMMUNOGLOBULIN SUPERFAMILY. BTN/MOG 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 onward for an unknown reason.
CC -----
CC This SWISS-PROT entry is copyright. It is produced through a collaboration between the Swiss Institute of Bioinformatics and the EMBL outstation - the European Bioinformatics Institute. There are no restrictions on its use by non-profit institutions as long as its content is in no way modified and this statement is not removed. Usage by and for commercial entities requires a license agreement (See <http://www.isb-sib.ch/announce/> or send an email to license@isb-sib.ch).
CC -----
DR EMBL; AF199028; AAC34739.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF289028; AAG01176.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AF216749; AAK16241.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; ABO14553; BAA31628.1; ALT_SEQ. [EMBL / GenBank / DDBJ] [CoCodingSequence]
DR EMBL; AX100595; CAC36465.1; -. [EMBL / GenBank / DDBJ] [CoCodingSequence]
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 -> ESWNLLL (IN ISOFORM 2).
SQ SEQUENCE 302 AA: 33349 MW: 647934E21B55E34A CRC64:
MRLGSPGLF LLFSSLRADT QKEVFRAMVG SDVELSCAP EGSRFDLNDV VVWQTSSESK
TVTYTHIPQN SSLENVDSRY RNRLAHMSPG MLRGDFSLRL FNVTQPDQEKF FHCLVLSQL
GFEQVLSVETV TLHVAANFSV PVWSAPHPS QDELTTCTS INGYVRPNVY WINKNTNSLL
DQLQNDTFV LMHGRGLYDV SVLRVIARTPS VNIGCCEVN LQQNLTGVS QTGNDIGERD
KITEENPVSTG EKNAATWSIL AVLCLVVVA VAIIGHVCRD CLQHSYAGAW AVSPTELTG
//

NiceProt View of SWISS-PROT: P01588

[Printer-friendly view](#)
[Quick BlastP search](#)
[\[General Information\]](#)
[\[ID\]](#)
[\[Accession\]](#)
[\[Comments\]](#)
[\[Cross-references\]](#)
[\[Keywords\]](#)
[\[Features\]](#)
[\[Sequence\]](#)
[\[Tools\]](#)

General information about the entry

Entry name EPO_HUMAN
 Primary accession number P01588
 Secondary accession numbers Q9UBHA0 Q9UEZ5 Q9UDZ0
 Entered in SWISS-PROT in Release 01, July 1986
 Sequence was last modified in Release 01, July 1986
 Annotations were last modified in Release 40, October 2000

Name and origin of the protein

Protein name ERYTHROPOIETIN [Precursor]
 Synonyms None

Gene name EPO

From Homo sapiens (Human) [TaxID: 9606]

Taxonomy Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominoidea; Homo.

References

- [1] SEQUENCE FROM NUCLEIC ACID.
 MEDLINE=85137899, PubMed=3838366, [NCBI, ExPASy, EBI, Israel, Japan]
 Jacobs K., Shoemaker C., Rutherford R., Nell S.D., Kaufman R.J., Mufson A., Seeger J., Jones S.S., Hewick R., Fritsch E.F., Kawakita M., Shimizu T., Miyake T.;
 "Isolation and characterization of genomic and cDNA clones of human erythropoietin.",
Nature 313:806-810(1985).
- [2] SEQUENCE FROM NUCLEIC ACID.
 MEDLINE=86067948, PubMed=3835178, [NCBI, ExPASy, EBI, Israel, Japan]
 Lin F.-K., Sugge S., Lin C.-H., Browne J.K., Smalling R., Egrie J.C., Chen K.K., Fox G.M., Martin F., Stabinsky Z., Badrawi S.M., Lai P.-H., Goldwasser E.;
 "Cloning and expression of the human erythropoietin gene.",
Proc Natl Acad Sci U S A 82:7580-7584(1985).
- [3] SEQUENCE FROM NUCLEIC ACID.
 MEDLINE=99011118, PubMed=9799793, [NCBI, ExPASy, EBI, Israel, Japan]
 Goeckner G., Scherer S., Schattauer R., Bongard A., Weber J., Tsu L.-C., Rosenthal A.;
 "Large-scale sequencing of two regions in human chromosome 7q22: analysis of 650 kb of genomic sequence around the EPO and CUTL1 loci reveals 17 genes.",
Genome Res 8:1060-1073(1998).
- [4] SEQUENCE FROM NUCLEIC ACID.
 Rupert J.L., Hochachka P.W.;
 "Erythropoietin gene sequence in the Quechua, a high altitude native population.",
 Submitted (NOV-1999) to the EMBL/GenBank/DBJ databases.
- [5] SEQUENCE OF P8_193 FROM NUCLEIC ACID, AND VARIANTS HEPATOCELLULAR CARCINOMA.
 MEDLINE=93384593, PubMed=8396923, [NCBI, ExPASy, EBI, Israel, Japan]
 Funakoshi A., Muta H., Baba T., Shimizu S.;
 "Gene expression of mutant erythropoietin in hepatocellular carcinoma.",
Biochem Biophys Res Commun 195:717-722(1993).
- [6] SEQUENCE OF P8_193, AND DISULFIDE BONDS.
 TISSUE=Uterus.
 MEDLINE=86140080, PubMed=3949763, [NCBI, ExPASy, EBI, Israel, Japan]
 Lai P.H., Everett R., Wang F.E., Arakawa T., Goldwasser E.;
 "Structural characterization of human erythropoietin.",
J Biol Chem 261:3116-3121(1986).
- [7] PRELIMINARY SEQUENCE OF P8_57.
 MEDLINE=84135751, PubMed=6698989, [NCBI, ExPASy, EBI, Israel, Japan]
 Yanagawa S., Hirade K., Ohnata H., Sasaki K., Cheha H., Ueda M., Goto M.;
 "Isolation of human erythropoietin with monoclonal antibodies.",
J Biol Chem 259:2707-2710(1984).
- [8] STRUCTURE OF CARBOHYDRATES.
 MEDLINE=88153657, PubMed=3346214, [NCBI, ExPASy, EBI, Israel, Japan]
 Takeuchi M., Takasaki S., Miyazaki H., Kato T., Hoshi S., Kubota N., Kobata A.;
 "Comparative study of the asparagine-linked sugar chains of human erythropoietins purified from urine and the culture medium of recombinant Chinese hamster ovary cells.",
J Biol Chem 263:3637-3663(1988).
- [9] STRUCTURE OF CARBOHYDRATES.
 MEDLINE=89118279, PubMed=3219367, [NCBI, ExPASy, EBI, Israel, Japan]
 Sasaki H., Ochi N., Dell A., Fukuda M.;
 "Site-specific glycosylation of human recombinant erythropoietin: analysis of glycopeptides or peptides at each glycosylation site by fast atom bombardment mass spectrometry.",
Biochemistry 27:8618-8626(1988).
- [10] STRUCTURE OF CARBOHYDRATES.
 MEDLINE=92314463, PubMed=1820196, [NCBI, ExPASy, EBI, Israel, Japan]
 Takeuchi M., Kobata A.;
 "Structures and functional roles of the sugar chains of human erythropoietins.",
Glycobiology 1:337-346(1991).

Comments

- **FUNCTION** ERYTHROPOIETIN IS THE PRINCIPAL HORMONE INVOLVED IN THE REGULATION OF ERTHROCYTE DIFFERENTIATION AND THE MAINTENANCE OF A PHYSIOLOGICAL LEVEL OF CIRCULATING ERTHROCYTE MASS.
- **SUBCELLULAR LOCATION** SECRETED
- **TISSUE SPECIFICITY** PRODUCED BY KIDNEY OR LIVER OF ADULT MAMMALS AND BY LIVER OF FETAL OR NEONATAL MAMMALS
- **PHARMACEUTICAL** AVAILABLE UNDER THE NAMES EPOGEN (AMGEN) AND PROCRIT (ORTHO BIOTECH)
- **SIMILARITY** BELONGS TO THE EPO / TPO FAMILY
- **DATABASE NAME** R&D Systems' cytokine source book; WWW=http://www.rndsystems.com/cyt_cat/epo.htm.

Copyright

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

Cross-references

EMBL	X02158; CAA26095.1; -	[EMBL / GenBank / DDBJ / CoDingSequence]
	X02157; CAA26094.1; -	[EMBL / GenBank / DDBJ / CoDingSequence]
	M11319; AAA52400.1; -	[EMBL / GenBank / DDBJ / CoDingSequence]
	AF05356; AAC78791.1; -	[EMBL / GenBank / DDBJ / CoDingSequence]
	AF202308; AACF23132.1; JOINED	[EMBL / GenBank / DDBJ / CoDingSequence]
	AF202307; AACF23132.1; JOINED	[EMBL / GenBank / DDBJ / CoDingSequence]
	AF202310; AACF23133.1; JOINED	[EMBL / GenBank / DDBJ / CoDingSequence]
	AF202311; AACF17572.1; -	[EMBL / GenBank / DDBJ / CoDingSequence]
	AF202314; AACF23134.1; -	[EMBL / GenBank / DDBJ / CoDingSequence]
	AF202313; AACF23134.1; JOINED	[EMBL / GenBank / DDBJ / CoDingSequence]
	S65458; AAD13964.1; -	[EMBL / GenBank / DDBJ / CoDingSequence]
PIR	A0185; ZUHU	
	A25384; A25384	
	A24744; A24744	
	A22210; A22210	
GlycoSuiteDB	P01588; -	
	MIM 133170 [NCBI / EBI]	
InterPro	EPO	
	IPR001323; EPO_TPO	
	IPR003013; Erythropoietin	
	Graphical view of domain structure	
	Pfam PF00758; EPO_TPO_1	
	PRINTS PR00272; ERYTHROPTN	
	PROSITE PS00817; EPO_TPO_1	
	ProDom ProDom / List of seq sharing at least 1 domain	
	BLOCKS P01588	
	DOMO P01588	
ProtoMap	ProtoMap P01588	
	PRESAGE P01588	
DIP	DIP P01588	
	SWISS-2DPAGE GET REGION ON 2D PAGE	
Keywords	Erythrocyte maturation, Glycoprotein, Hormone, Signal, Pharmaceutical	
	Feature	
		Feature aligner
		Feature table viewer
		Sequence information
		Length: 193 AA [This is the length of the unprocessed precursor]
		Molecular weight: 21306 Da [This is the MW of the unprocessed precursor]
		CRC64: C91F0E4C26A52033 [This is a checksum on the sequence]
		MGVHECPANL VLLSLSSLSP LGLPVLVAPP RLICDSDVLE SYLLEAKEAR NITTGCAEHC
		70 80 90 100 110 120
P01588		SLNENITVPD TKVNFTYAWKR NEVQQQAVEW WQQLALLSEK VLRLQALLWN SQQPWEPLQL
		130 140 150 160 170 180
		HVDKAVSGRL SLTTLLRALG AKREASIPPD AASAAPLRTI TADTFKLFLR VYSNFLRGKL
		190
		KLYTGAEACT GDR

[View entry in original SWISS-PROT format](#)

[View entry in raw text format \(no links\)](#)

[Report form for errors/updates in this SWISS-PROT entry](#)

Direct BLAST submission at [EMBnet-CH and CSCS \(Switzerland\)](#)

ScanProsite, [ProfileScan](#)

Feature table [viewer](#) (Java)

Direct BLAST submission at [NCBI \(Bethesda, USA\)](#)

Tools, [Sequence analysis tools](#): [ProParam](#), [ProtScale](#), [Compute pI/Mw](#), [PeptideMass](#), [Dotter](#) (Java)

Search the [SWISS-MODEL Repository](#)

TrEMBL: example

TrEMBL: Q9UDZ0

ID Q9UDZ0 PRELIMINARY; PRT; 136 AA.

AC Q9UDZ0;

DT 01-MAY-2000 (TrEMBLrel. 13, Created)

DT 01-MAY-2000 (TrEMBLrel. 13, Last sequence update)

DT 01-JUN-2000 (TrEMBLrel. 14, Last annotation update)

DE ERYTHROPOIETIN PROTEIN (FRAGMENT).

GN ERYTHROPOIETIN.

OS Homo sapiens (Human).

OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;

OC Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

RN [1]

RP SEQUENCE FROM N.A.

RX MEDLINE; 93384593. [[NCBI](#), [ExPASy](#), [Israel](#), [Japan](#)]

RA Funakoshi A., Muta H., Baba T., Shimizu S.;

RT "Gene expression of mutant erythropoietin in hepatocellular carcinoma.";

RL Biochem. Biophys. Res. Commun. 195(2):717-722 (1993).

DR EMBL; S65458; AAD13964.1; -. [[EMBL](#) / [GenBank](#) / [DDBJ](#)] [[CoCodingSequence](#)]

DR [INTERPRO](#); IPR001323; -.

DR [INTERPRO](#); IPR003013; -.

DR [PFAM](#); PF00758; EPO_TPO; 1.

DR [PRINTS](#); PRO0272; ERYTHROPTN.

DR [PROTOMAP](#); Q9UDZ0.

DR [PRESAGE](#); Q9UDZ0.

DR [SWISS-2DPAGE](#); GET REGION ON 2D PAGE.

FT NON_TER 1 1

SQ SEQUENCE 136 AA; 15173 MW; BCB9B1F0D8190AB3 CRC64;
EHCSLNENIT VPDTKVNFYA WKRMEVGQQA VEVWQGLALL SEAVLRGQAL LVNSSQPWE
LQLHVDKAVS GLRNFTLLR ALGAQKEAIS PQDAASAAPL RTITADTFRK LFRVYSNFLR
GKLKLYTGEA CRTGDR

//

Original TrEMBL entry which has been integrated into the SWISS-PROT EPO_HUMAN entry and thus which is not found in TrEMBL anymore.

NiceProt View of SWISS-PROT: [P01588](#)

[Printer-friendly view](#)

Quick BlastP search

[General](#) | [Name and origin](#) | [References](#) | [Comments](#) | [Cross-references](#) | [Keywords](#) | [Features](#) | [Sequence](#) | [Tools](#)

General information about the entry

Entry name	EPO_HUMAN
Primary accession number	P01588
Secondary accession numbers	Q9UHA0 Q9UEZ5 Q9UDZ0
Entered in SWISS-PROT in	Release 01, July 1986
Sequence was last modified in	Release 01, July 1986
Annotations were last modified in	Release 40, October 2000

Name and origin of the protein	
Protein name	ERYTHROPOEITIN [Precursor]
Synonyms	None
Gene name	EPO
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Bukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominoidea; Homo.

References

- [1] SEQUENCE FROM NUCLEIC ACID
MEDLINE=95137899, PubMed=38382366, [NCBI ExPASy, EBI Israel, Japan]
Jacobs K, Sheenaker C, Rudererdorf R, Neff S.D., Kaufman R.J., Mufson A., Seehra J., Jones S.S., Hewick R., Fritsch E.F., Kawakita M., Shimizu T., Miyake T.;
"Isolation and characterization of genomic and cDNA clones of human erythropoietin",
Nature 313:806-810(1985).

[2] SEQUENCE FROM NUCLEIC ACID
MEDLINE=86067948, PubMed=3865178, [NCBI ExPASy, EBI Israel, Japan]
Lei P.-K., Sugge S., Lin C.-H., Browne K., Smalling R., Egrie J.C., Chen K.K., Fox G.M., Martin F., Stabinsky Z., Badravi S.M., Lai P.-H., Goldwasser E.;
"Cloning and expression of the human erythropoietin gene",
Proc Natl Acad Sci U.S.A. 82:7580-7584(1985)

[3] SEQUENCE FROM NUCLEIC ACID
MEDLINE=99018118, PubMed=9799793, [NCBI ExPASy, EBI Israel, Japan]
Gloeckner G., Scherer S., Schatheyev R., Boright A., Weber J., Tsui L.-C., Rosenthal A.;
"Large-scale sequencing of two regions in human chromosome 7q22: analysis of 650 kb of genomic sequence around the EPO and C UTIL loci reveals 17 genes",
Genome Res. 8:1060-1073(1998)

[4] SEQUENCE FROM NUCLEIC ACID
Rupert JL, Horchakova P.W.;
"Erythropoietin gene sequence in the Quechua, a high altitude native population",
Submitted (NOV-1999) to the EMBL/GenBank/DDBJ databases

[5] SEQUENCE OF 52-123 FROM NUCLEIC ACID, AND VARIANTS HEPATOCELLULAR CARCINOMA
MEDLINE=93384593, PubMed=8396923, [NCBI ExPASy, EBI Israel, Japan]
Funakoshi A., Mata H., Baba T., Shimizu S.;
"Gene expression of mutant erythropoietin in hepato-cellular carcinoma",
Biochem Biophys Res Commun 193:717-722(1993)

[6] SEQUENCE OF 28-193, AND DISULFIDE BONDS:
TISSUE=Urine.
MEDLINE=86140080, PubMed=3949763, [NCBI ExPASy, EBI Israel, Japan]
Lei P.-K., Everett R., Wang F.F., Arakawa T., Goldwasser E.;
"Structural characterization of human erythropoietin",
J Biol Chem 261:3116-3121(1986)

[7] PRELIMINARY SEQUENCE OF 28-57
MEDLINE=94135751, PubMed=6698989, [NCBI ExPASy, EBI Israel, Japan]
Yanagawa S., Hirade K., Ohnata H., Sasaki R., Chiba H., Ueda M., Goto M.;
"Isolation of human erythropoietin with monoclonal antibodies",
J Biol Chem 259:2707-2710(1984)

[8] STRUCTURE OF CARBOHYDRATES
MEDLINE=83153657, PubMed=3346214, [NCBI ExPASy, EBI Israel, Japan]
Takeuchi M., Takasaki S., Miyazaki H., Kato T., Hoshii S., Kochibe N., Kobata A.;
"Comparative study of the asparagine-linked sugar chains of human erythropoietins purified from urine and the culture medium of recombinant Chinese hamster ovary cells",
J Biol Chem 263:3657-3663(1988)

[9] STRUCTURE OF CARBOHYDRATES
MEDLINE=99118279, PubMed=3219367, [NCBI ExPASy, EBI Israel, Japan]
Sasaki H., Ochiai N., Dell A., Fukuda M.;
"Site-specific glycosylation of human recombinant erythropoietin: analysis of glycopeptides or peptides at each glycosylation site by fast atom bombardment mass spectrometry",
Biochemistry 27:8618-8626(1988)

[10] STRUCTURE OF CARBOHYDRATES
MEDLINE=92314463, PubMed=1820196, [NCBI ExPASy, EBI Israel, Japan]
Takeuchi M., Kobata A.;
"Structures and functional roles of the sugar chains of human erythropoietins",
Glycobiology 1:337-346(1991).

Comment

- **FUNCTION** ERYTHROPOEITIN IS THE PRINCIPAL HORMONE INVOLVED IN THE REGULATION OF ERYTHROCYTE DIFFERENTIATION AND THE MAINTENANCE OF A PHYSIOLOGICAL LEVEL OF CIRCULATING ERYTHROCYTE MASS.
 - **SUBCELLULAR LOCATION** SECRETED.
 - **TISSUE SPECIFICITY** PRODUCED BY KIDNEY OR LIVER OF ADULT MAMMALS AND BY LIVER OF FETAL OR NEONATAL MAMMALS.
 - **PHARMACEUTICAL** AVAILABLE UNDER THE NAMES EPOGEN (AMGEN) AND PROCRIT (ORTHO BIOTECH).
 - **SIMILARITY** BELONGS TO THE EPO / TPO FAMILY.
 - **DATABASE NAME**=R&D Systems cytokine source book, WWW=http://www.mdsystems.com/cyt_cat/epo.htm.

Copyright

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

Cross-references

X02158, CAA26095_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
X02157, CAA26094_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
M11319, AAA52400_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF053356, AAC78791_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202308, AAF23132_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202306, AAF23132_1, JOINED	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
EMBL AF202307, AAF23132_1, JOINED	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202310, AAF23132_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202309, AAF23132_1, JOINED	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202311, AAIF17572_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202314, AAF23134_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202312, AAF23134_1, JOINED	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
AF202313, AAF23134_1, JOINED	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
SE5458, AAD13964_1, -	[EMBL / GenBank / DDBJ] [CoLoringSequence]			
PIR A01855, ZUHU				
A25384, A25384				
A24744, A24744				
A22210, A22210				
GlycoSuteDB P01588_-,				
MIM 133170 [NCBI / EBI]				
GeneCards EPO				
InterPro IPR001323_EPO_TPO				
InterPro IPR030135_Erythroin				
Pfam Graphical view of domain structure.				
PRINTS PF00758_EPO_TPO_1				
PROSITE PR00272_ERYTTHROIN				
ProDom (Domain structure / List of seq. sharing at least 1 domain)				
BLOCKS P01588				
DOMO P01588				
ProtoMap P01588				
PRESAGE P01588				
DIP P01588				
SWISS-2DPAGE GET REGION ON 2D PAGE				
Keywords				
Erythrocite maturation, Glycoprotein, Hormone, Signal, Pharmaceutical.				
Features				
Key	From	To	Length	Description
SIGNAL	1	27	27	
CHAIN	28	193	166	ERYTHROPOIETIN.
PROPEP	190	193	4	May BE REMOVED IN PROCESSED PROTEIN.
DISULFID	34	188	155	
DISULFID	56	60	5	
CARBOHYD	51	51	1	N-LINKED (GLCNAC...).
CARBOHYD	65	65	1	N-LINKED (GLCNAC...).
CARBOHYD	110	110	1	N-LINKED (GLCNAC...).
CARBOHYD	153	153	1	O-LINKED (GALNAc...).
VARIANT	131	132	2	SL --> NF (IN AN HEPATOCELLULAR CARCINOMA). /FTId=VAR_009870.
VARIANT	149	149	1	P --> Q (IN AN HEPATOCELLULAR CARCINOMA). /FTId=VAR_009870.
CONFLICT	40	40	1	E --> Q (IN REF. 1: CAA26095) .
CONFLICT	85	85	1	Q --> QQ (IN REF. 5) .
CONFLICT	140	140	1	G --> R (IN REF. 1: CAA26095) .
Sequence information				
Length 193 AA [This is the length of the unprocessed precursor]	Molecular weight 21306 Da [This is the MW of the unprocessed precursor]	Sequence		
10 20 30 40 50 60				
MGVHECPAW LLLSLLLSLP LGLPVLPGLAPP RLICDSWPLV SYLLEKAEKA NITTGCAERC				
70 80 90 100 110 120				
SLNLENITVPP TRUNFYNAWR MEUVQQAAVEY WQQLALLSEA VLRLQGALLVNV S2QPEWPLQL				
130 140 150 160 170 180				
HVDFAVSGRL SLTTLLRALQ AQKEAISPPD AASAAAPLRTI TADTFRKLFV VYSNFLRGKL				
190				
KLYVGEACRT GDR				

ISOTOPIC COMPOSITION OF CHICKEN PROTEIN

[View entry in raw text format \(no
links\)](#)

[Report form for error/updates in this SWISS-PROT entry](#)



 Direct BLAST submission at [EMBnet-CH](#) and CSC
(Switzerland)



 ScanProsite ProfileScan



3



Direct BLAST submission at [NCBI \(Bethesda, US\)](#)



Sequence analysis tools: [ProtParam](#), [ProtScale](#), [Compute pI/Mw](#)



S. A. S. CHENG, M. O. HUSSAIN

SWISS-PROT and the cross-references (X-ref)

- SWISS-PROT was the 1st database with X-ref.;
- Explicitly X-referenced to 36 databases;
X-ref to DNA (EMBL/GenBank/DDBJ), 3D-structure (PDB), literature (Medline), genomic (MIM, MGD, FlyBase, SGD, SubtiList, etc.), 2D-gel (SWISS-2DPAGE), specialized db (PROSITE, TRANSFAC);
- Implicitly X-referenced to 17 additional db added by the ExPASy servers on the WWW (i.e.: GeneCards, PRODOM, HUGE, etc.)

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

**Domains, functional sites,
protein families**

PROSITE

InterPro

Pfam

PRINTS

SMART

Mendel-GFDb

2D and 3D Structural dbs

HSSP

PDB

PTM

CarbBank

GlycoSuiteDB

2D-gel protein databases

SWISS-2DPAGE

ECO2DBASE

HSC-2DPAGE

Aarhus and Ghent

MAIZE-2DPAGE

**Human diseases
MIM**

Protein-specific dbs

GCRDb

MEROPS

REBASE

TRANSFAC

Organism-spec. dbs

DictyDb

EcoGene

FlyBase

HIV

MaizeDB

MGD

SGD

StyGene

SubtiList

TIGR

TubercuList

WormPep

Zebrafish

SWISS-PROT

Nucleotide sequence db
EMBL, GeneBank, DDBJ

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

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.

[iProClass](#), 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](#)



The screenshot shows a search results page for the 'S'-Nucleotidase Family Members' database. At the top, there is a legend for motif signatures: POM0010 (red), POM0098 (orange), POM0099 (yellow), POM0140 (blue), and POM0072 (green). Below the legend, five protein sequences are listed with their accession numbers: A2679, B16059, B16054, B16052, and V16059. Each sequence is represented by a grey bar with colored dots indicating motif locations. To the right of the sequences is a 3D ribbon diagram of a protein structure.



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)

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.; Mufson, A.; Seehra, J.; Jones, S.S.;
 Hewick, R.; Fritsch, E.F.; Kawakita, 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:85137899
 #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.; Smalling,
 R.; Egrie, J.C.; Chen, K.K.; Fox, G.M.; Martin, F.;
 Stabinsky, Z.; Badravi, 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; PID: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.; Ohnoto, 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
 51,65,110
 #disulfide bonds #status experimental
 #binding site carbohydrate (Asn) (covalent)
 #status experimental
 153
 #binding site carbohydrate (Ser) (covalent)
 #status experimental
 SUMMARY #length 193 #molecular_weight 21306
 SEQUENCE
 5 10 15 20 25 30
 1 M G V H E C P A W L W L L S L L S L P L G 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 V E V
 91 W Q G L A L S 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 R A L G A Q K E A I S P P D
 151 A 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

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

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
 51,65,110

#disulfide bonds #status experimental
 #binding site carbohydrate (Asn) (covalent)

#status experimental

#binding site carbohydrate (Ser) (covalent)

#status experimental

153

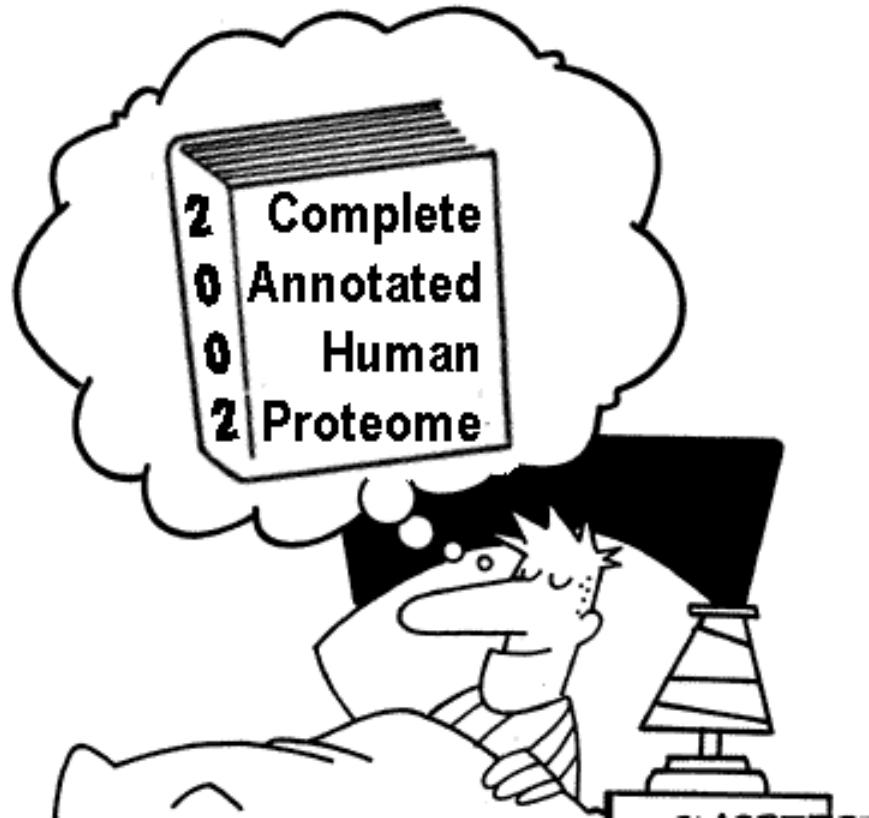
#status experimental

GenPept (translation of GenBank)

- GenPept is a protein database translated from the last release of GenBank (+ journal scans)
- The current release has > 1 million entries
- In contrast to TrEMBL, keeps all protein sequences including small fragments (< 8 aa), immunoglobulins....
- Redundancy: > 20 entries for human EPO



When Amos dreams...



Database 5: 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 »)



Database 4: 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 ⊕ EPO_HUMAN			2.41	APPRLICDSRVLERYLLEAKEAENVTMGCSEHCSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQG					
2 ⊕ EPO_RAT			2.61	APPRLICDSRVLERYILEAKEAENVTMGCAGPRLSENITVPDTKVNFYAWKRMEVEEQAIENVWQG					
3 ⊕ EPO_FELCA			2.99	APPRLICDSRVLERYILEAREAAENATMGCAEGCSFSENITVPDTKVNFYAWKRMEVQQQALEEVWQG					
8 Consensus			8.01	APPRLICDSRVLERYILEAKEAENVTMGCAGCSSLNENITVPDTKVNFYAWKRMEVGQQAVEVWQG					
1 PROSITE				-----	-----	-----	-----	-----	-----

- 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.



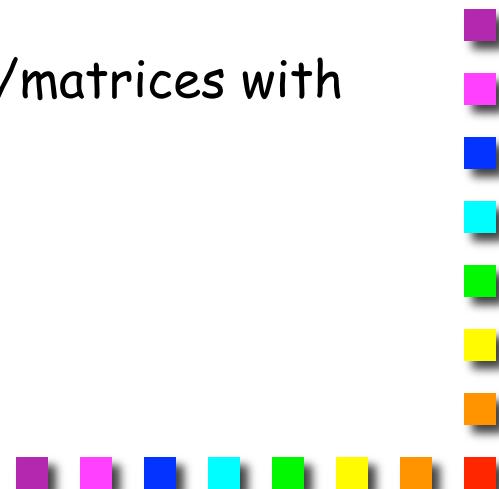
Protein domain/family db

- 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)



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	
<ul style="list-style-type: none">• 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

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](#))

SWISS-PROT

'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](#))

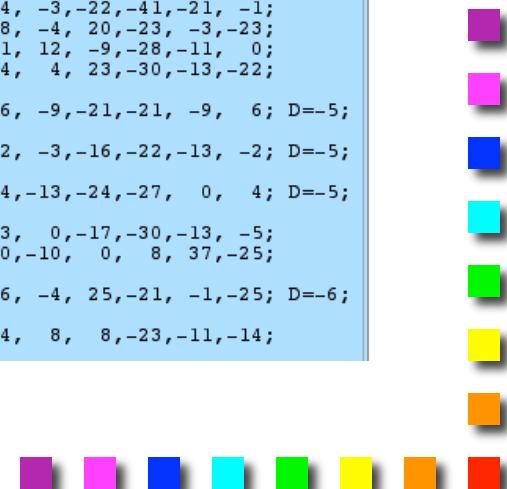
Retrieve an alignment of SWISS-PROT true positive hits:

[[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

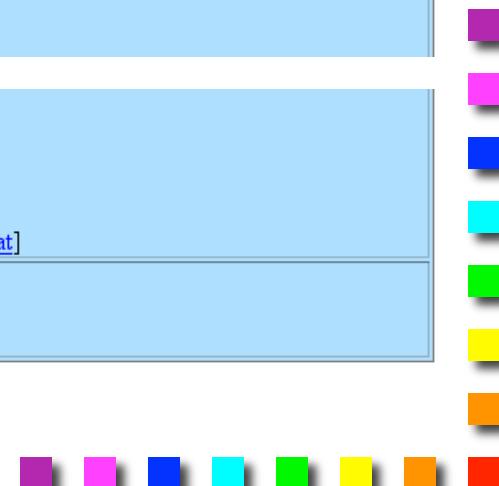
- SWISS-PROT release number: **40.17**, total number of sequence entries in that release: **108489**.
- Total number of hits in SWISS-PROT: **68 hits in 59 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **68 hits in 59 different sequences**
- 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: **Archaeabacteria, 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_RHOMP (Q30477), DNAB_SVNV2 (Q55418), DDT_DVRBP (Q9Y2F4),
	 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  1AM2 ; Detailed view	

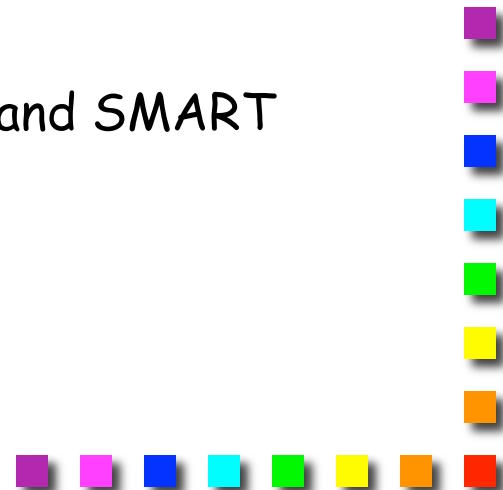


Protein domain/family db

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

Interpro

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



InterPro: www.ebi.ac.uk/interpro

 **EMBL-EBI**
European Bioinformatics Institute

Get Nucleotide sequences for Go ? Site search Go ?

EBI Home About EBI Research Services Toolbox Databases Downloads Submissions
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



A vertical column of colored squares (purple, magenta, blue, cyan, green, yellow, orange, red) is positioned on the right side of the page.

Some statistics

- 15 most common domains for *H. sapiens* (Incomplete)

■	InterPro	Matches(Proteins matched)	Name
■	IPR000822	30034(1093)	Zn-finger, C2H2 type
■	IPR003006	2631(1032)	Immunoglobulin/major histocompatibility complex
■	IPR000561	4985(471)	EGF-like domain
■	IPR001841	1356(458)	Zn-finger, RING
■	IPR001356	2542(417)	Homeobox
■	IPR001849	1236(405)	Pleckstrin-like
■	IPR000504	2046(400)	RNA-binding region RNP-1 (RNA recognition motif)
■	IPR001452	2562(394)	SH3 domain
■	IPR002048	2518(392)	Calcium-binding EF-hand
■	IPR003961	2199(300)	Fibronectin, type III
■	IPR001478	1398(280)	PDZ/DHR/GLGF domain
■	IPR005225	261(261)	Small GTP-binding protein domain
■	IPR000210	583(236)	BTB/POZ domain
■	IPR001092	713(226)	Basic helix-loop-helix dimerization domain bHLH
■	IPR002126	5168(226)	Cadherin

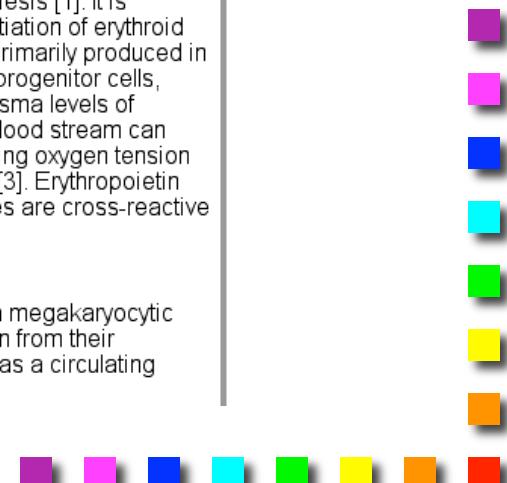


InterPro example

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 	<p>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].</p> <p>Thrombopoietin (TPO), a glycoprotein, is the <u>mammalian</u>0 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.</p>



InterPro example

Examples	<ul style="list-style-type: none">P49745 TPO_RATP33709 EPO_SHEEPP33708 EPO_FELCA
	View examples
References	<ol style="list-style-type: none">Shoemaker C.B., Mitsock L.D. <i>Murine erythropoietin gene - Cloning, expression , and human gene homology.</i> Mol. Cell. Biol. 6: 849-858(1986). [MEDLINE:87039105]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]Lin F.K., Lin C.H., Lai P.H., Browne J.K., Egrie J.C., Smallling 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]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 <u>P01588</u>	IPR001323 PS00817 EPO_TPO
	IPR001323 PF00758 EPO_TPO
	IPR003013 PR00272 ERYTHROPTN
SWISS-PROT EPO_MOUSE <u>P07321</u>	IPR001323 PS00817 EPO_TPO
	IPR001323 PF00758 EPO_TPO
	IPR003013 PR00272 ERYTHROPTN
SWISS-PROT EPO_MACFA <u>P07865</u>	IPR001323 PS00817 EPO_TPO
	IPR001323 PF00758 EPO_TPO
	IPR003013 PR00272 ERYTHROPTN
SWISS-PROT EPO_RAT <u>P29676</u>	IPR001323 PS00817 EPO_TPO
	IPR001323 PF00758 EPO_TPO
	IPR003013 PR00272 ERYTHROPTN
SWISS-PROT EPO_CANFA <u>P33707</u>	IPR001323 PS00817 EPO_TPO
	IPR001323 PF00758 EPO_TPO

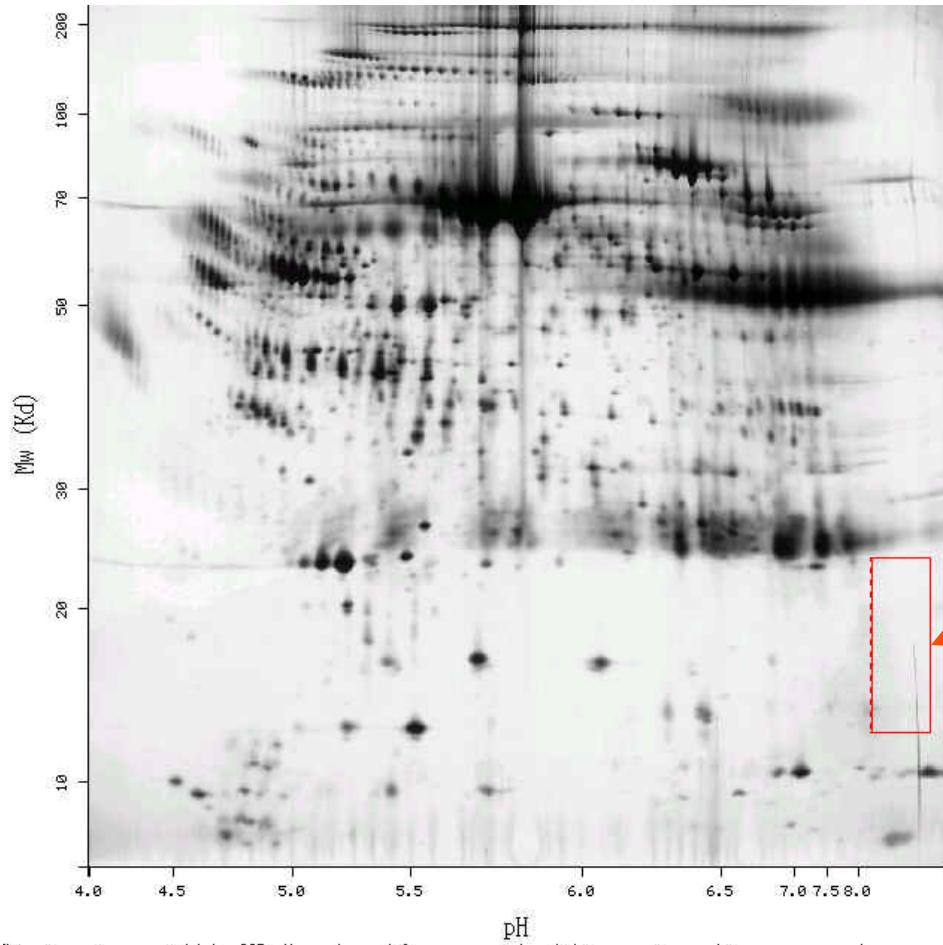


Databases 6: proteomics

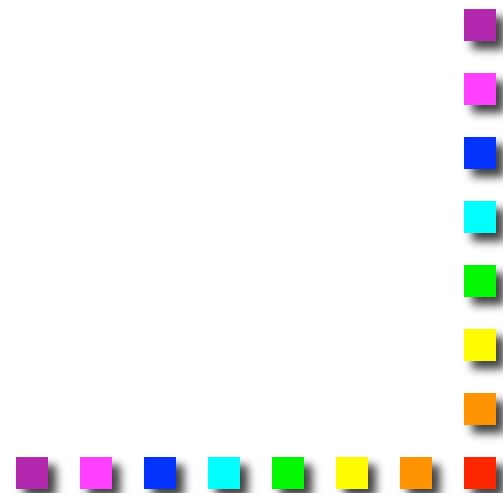
- Contain informations obtained by 2D-PAGE: master images of the gels and description of identified proteins
- Examples: SWISS-2DPAGE, ECO2DBASE, Maize-2DPAGE, Sub2D, Cyano2DBase, etc.
- Format: composed of image and text files
- Most 2D-PAGE databases are “federated” and use SWISS-PROT as a master index
- 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...



Databases 7: 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



PDB example 1eer

Databases 8: 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.



Databases 9: bibliographic

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



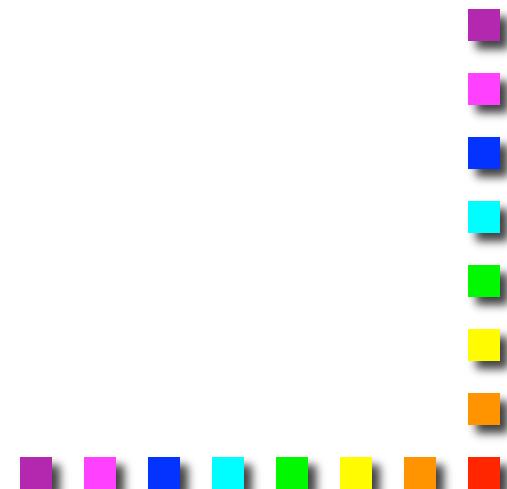
Medline

- 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)



Databases 10: 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.



Proliferation of databases

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



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)



Database 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)
-





Top Page

Query

Results

Projects

Views

Databanks

HELP

Reset

Quick Search |

All Entries

Query forms

 Standard Extended

► bookmark this
[link](#) to return to
your session

► [BookmarkLet](#) for:

- Protein Seq
- DNA/RNA Seq
- Structures

If you find problems
or have suggestions
please mail the
[SRS administrator](#)

show all collapse all **[+]** References all MEDLINE GO GOA

References - subsections

 all MEDLINE (Updates) MEDLINE (Main Release)**[+]** Sequence libraries - complete

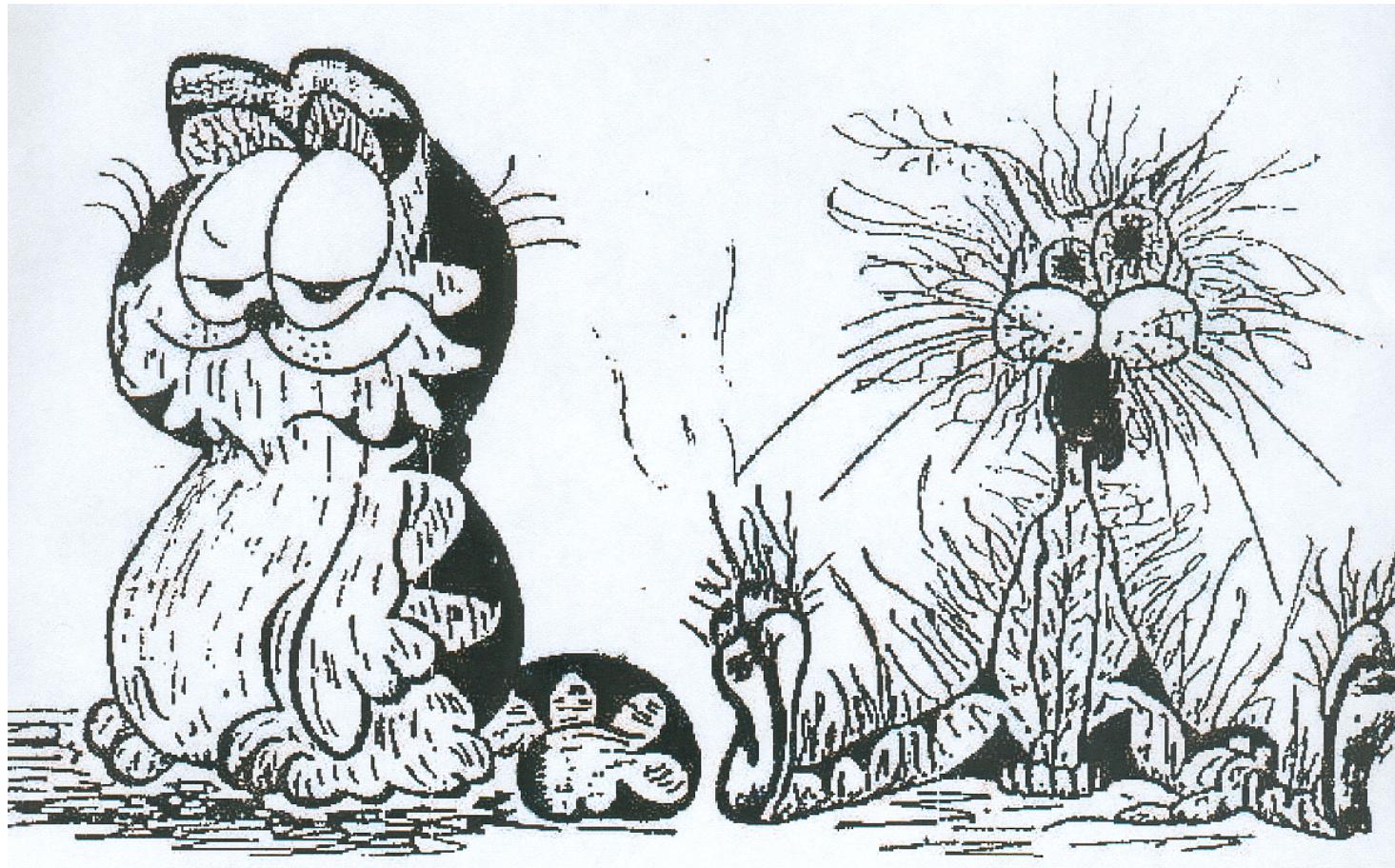
<input type="checkbox"/> all	<input type="checkbox"/> EMBL	<input type="checkbox"/> SWALL (SPTR)	<input type="checkbox"/> IPI	<input type="checkbox"/> RemTrEMBL
	<input type="checkbox"/> ENSEMBL HUMAN	<input type="checkbox"/> ENSEMBL MOUSE	<input type="checkbox"/> ENSEMBL FLY	<input type="checkbox"/> PATENT_PRT
	<input type="checkbox"/> JPO_PRT	<input type="checkbox"/> PATENT_DNA	<input type="checkbox"/> USPO_PRT	<input type="checkbox"/> ENSEMBL FISH
	<input type="checkbox"/> IMGT/LIGM-DB	<input type="checkbox"/> IMGTHLA	<input type="checkbox"/> MHCBN	

[+] Sequence libraries - subsections

<input type="checkbox"/> all	<input type="checkbox"/> EMBL (Release)	<input type="checkbox"/> EMBL (Updates)	<input type="checkbox"/> EMBL (WGS)	<input type="checkbox"/> EMBL (TPA)
	<input type="checkbox"/> REFSEQ	<input type="checkbox"/> REFSEQP	<input type="checkbox"/> SWISS-PROT	<input type="checkbox"/> SpTrEMBL
	<input type="checkbox"/> TrEMBL (Updates)			

[+] InterPro&Related**[+]** SeqRelated**[+]** TransFac**[+]** User Owned Databanks**[+]** Application Results**[+]** Protein3DStruct**[+]** Genome**[+]** Mapping**[+]** Mutations**[+]** Locus Specific Mutations**[+]** Metabolic Pathways**[+]** Others**[+]** SNP**[+]** EMBOSS DOCS**[+]** System

Before the introduction to databases...



After the introduction to databases...