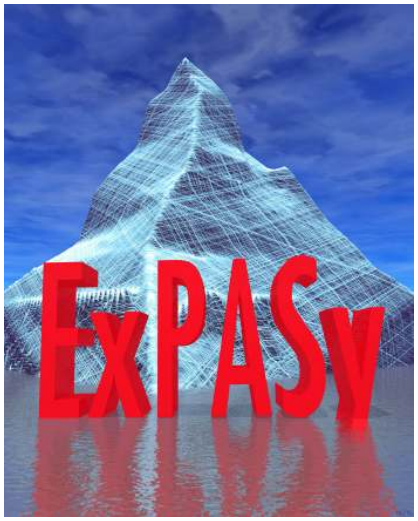


Protein sequence databases and proteomics tools developed at the

Swiss Institute of Bioinformatics

Elisabeth Gasteiger (Elisabeth.Gasteiger@isb-sib.ch)

Trieste, June 2006

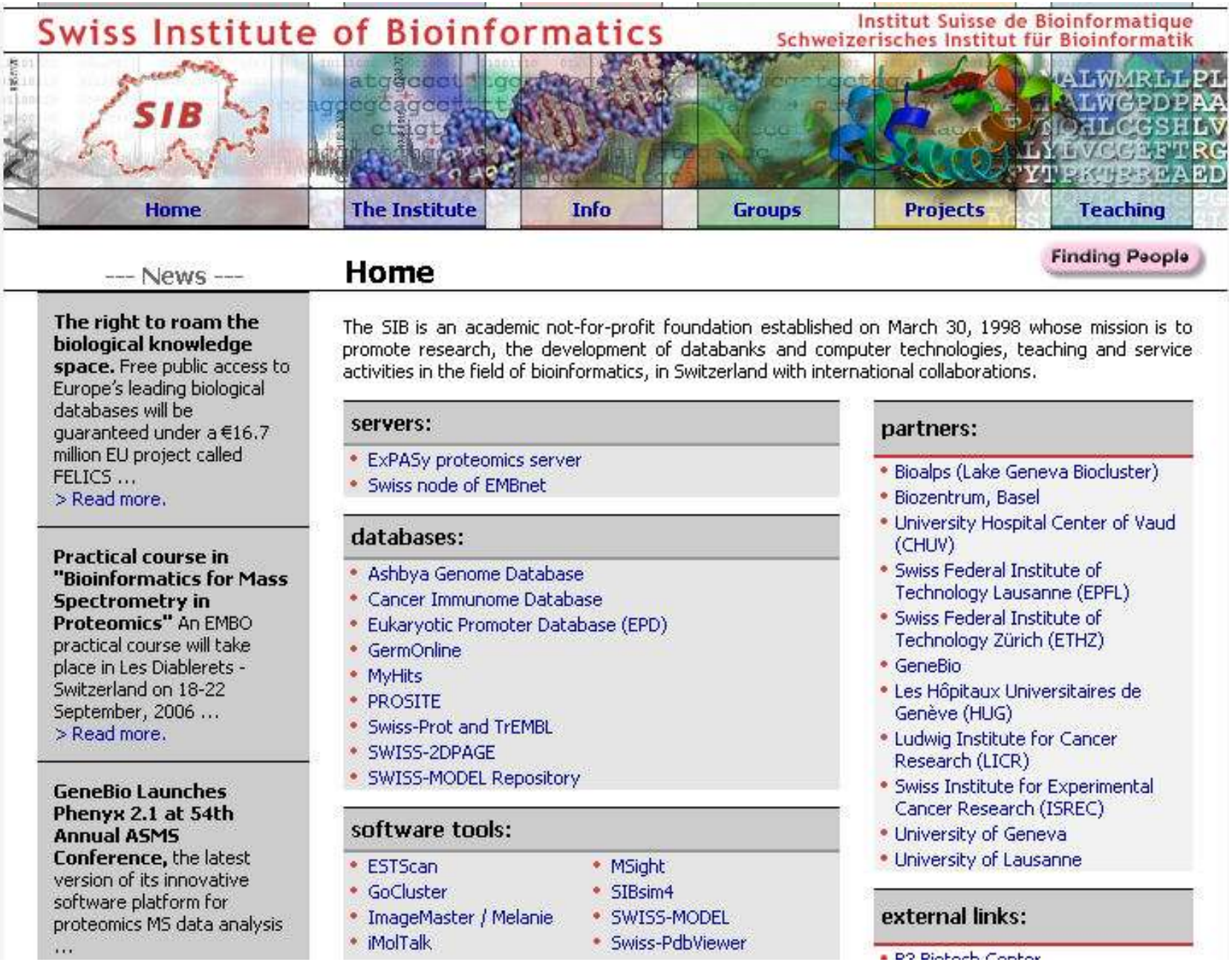
Outline

- The Swiss Institute of Bioinformatics
- What is UniProt?
- UniProt Knowledgebase: Swiss-Prot and TrEMBL
- UniRef and UniParc
- Non-redundancy, alternative splicing, post-translational modifications, SNPs in UniProtKB/Swiss-Prot
- Databases for protein function and domains: PROSITE, InterPro etc

Swiss Institute of Bioinformatics (SIB)

- Non-profit foundation created in 1998;
- Research, development, education and services in bioinformatics;
- Groups in Geneva,Lausanne and Basel;
- Federation of several groups (some of which existed and collaborated long before the foundation of the institute), about 170 researchers in 2006.

www.isb-sib.ch



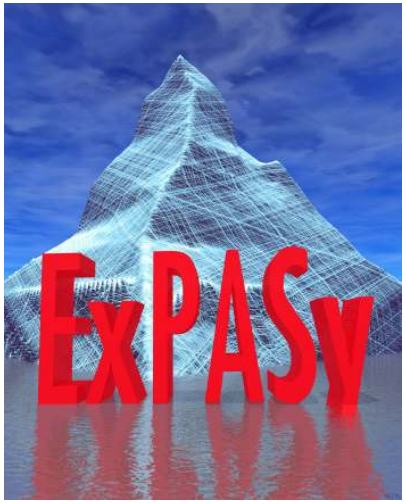
swissprot

The four missions of the SIB

- Development of databases and software tools;
- High-quality bioinformatics research program;
- Courses and seminars for the training of bioinformatics research scientists. This includes a master's degree in proteomics and bioinformatics, several weekly courses and a doctoral school (http://www.isb-sib.ch/teaching/intro.htm);
- Services to the Swiss Life Sciences community (EMBnet node).

Swiss Institute of Bioinformatics: 13 research and service groups

- Bioinformatics Core Facility
- Biological Information Modeling
- Computational Cancer Genomics - ISREC
- Computational Systems Biology
- EMBnet Facility
- Genome Bioinformatics
- Genome Systems Biology
- Molecular Modeling
- Protein Structure Bioinformatics
- Proteome Informatics (PIG)
- RNA regulatory networks
- Swiss-Prot
- Transcriptome Analysis
- Vital-IT

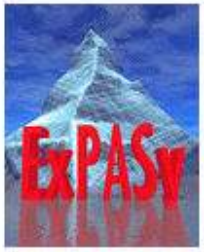


The ExPASy WWW server
www.expasy.org

- First molecular biology server on the Web (August 1993); ~500 million accesses since;
- Dedicated to proteomics:
  - Databases: **UniProtKB**, **PROSITE**, **Swiss-2DPAGE**, etc.;
  - Many 2D/MS protein identification/characterization and sequence analysis tools;
- Mirror sites in Australia, Brazil, Canada, China, Korea and USA: http://{au|br|ca|cn|us|www}.expasy.org

Site Map | Search ExPASy | Contact us

Search Swiss-Prot/TrEMBL for Go Clear



ExPASy Proteomics Server

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).

In-Silico Analysis of Proteins
Celebrating the 20th Anniversary of Swiss-Prot
- Register now!

[Announcements] [Job opening] [Mirror Sites]

Databases

- **UniProt Knowledgebase (Swiss-Prot and TrEMBL)** - Protein knowledgebase
- **PROSITE** - Protein families and domains
- **SWISS-2DPAGE** - Two-dimensional polyacrylamide gel electrophoresis
- **ENZYME** - Enzyme nomenclature
- **SWISS-MODEL Repository** - Automatically generated protein models
- Links to many other molecular biology databases

Tools and software packages

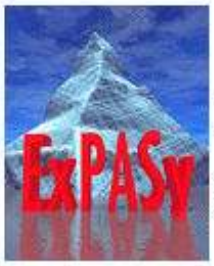
- **Proteomics and sequence analysis tools**
  - Proteomics
  - DNA -> Protein
  - Similarity searches (BLAST...)
  - Pattern and profile searches (ScanProsite...)
  - Post-translational modification and topology prediction
  - Primary structure analysis
  - Secondary and tertiary structure tools (Swiss-PdbViewer...)
  - Alignment and Phylogenetic analysis
- **ImageMaster / Melanie** - Software for 2-D PAGE analysis
- **MSight** - Mass Spectrometry Imager
- **Roche Applied Science's Biochemical Pathways**

Education and services

- **The ExPASy FTP server**
- **Swiss-Shop** - automatically obtain (by email) new sequence entries relevant to your field(s) of interest
- **Vital-IT** - The HPC Center for Life Sciences
- **e-Proxemis** - Proteomics-oriented Bioinformatics Training Portal

Documentation

- **What's New on ExPASy**
- **SWISS-FLASH** electronic bulletins
- **Swiss-Prot documents**
- **How to create HTML links to ExPASy**
- **Complete table of available documents**



ExPASy Life Science Directory

(formerly known as Amos' WWW links page)

http://www.expasy.org/links.html

Notes:

1) The URL for this page is http://www.expasy.org/links.html
2) If you would like to submit a specific link or to notify us of a modified link,please send us an email, but remember that we reserve the right to choose the links
want to vinclude !
3) Links to protein sequence, 3D structure and 2D-gel analytical tools are provided on ExPASy's Proteomics tools page.

Quick jump to the following topics:

Protein db | 3D structure db | 2D-PAGE & MS db | DNA/RNA db | Carbohydrates db | Organisms specific db | Human mutation db | Genes/proteins specific db | PTM
Phylogenetics db | Microarrays db | Patents | References | Dict., primers & nomenclat. | Biol. soft. & db catalogs | Gateways | Biol. journals & publishers | Biol. socie
Biocomputing servers | Biotech. companies | Bioinformatics companies | Misc. medical ref. sites | Misc. scientific ref. sites

---

Protein related databases

- Swiss-Prot - Swiss-Prot annotated protein sequence db
- Kabat - Kabat db of sequences of proteins of immunological interest
- PMD - Protein Mutant db

- InterPro - Integrated Resources of Proteins Domains and Functional Sites
- PROSITE - PROSITE dictionary of protein sites and patterns
- BLOCKS - BLOCKS db
- Pfam - Protein families db (HMM derived) [Mirrors at St. Louis (USA), Sanger Institute, UK, Karolinska Institutet (Sweden)]
- PRINTS - Protein Motif fingerprint db
- ProDom - Protein domain db (Automatically generated)
- PROTOMAP - An automatic hierarchical classification of Swiss-Prot proteins
- SBASE - SBASE domain db
- SMART - Simple Modular Architecture Research Tool
- STRING - Search Tool for the Retrieval of Interacting Genes/Proteins
- TIGRFAMs - TIGR protein families db

Protein sequence databases: History

- **1965** Atlas of Protein Sequence and Structure (65 proteins)
- **1980** ~ 80 genes fully sequenced
- **1982** EMBL
- **1983** Protein information ressource (PIR)
- **1986** Swiss-Prot is created by Amos Bairoch (~3900 proteins)
- **1996** TrEMBL (TRanslation of EMBL) (EBI).
  Complement to Swiss-Prot to cope with the enormous quantity of new sequences; contains all coding sequences not yet in Swiss-Prot

- **2002** UniProt consortium
- **2003** UniProt Release 1.0 public

- **2006** more than 3'000'000 protein sequences known !

1986-2006

Swiss-Prot: Alive and Kicking!

You are welcome to our 20th anniversary meeting in Brazil this year

http://www.swissprot20.org/

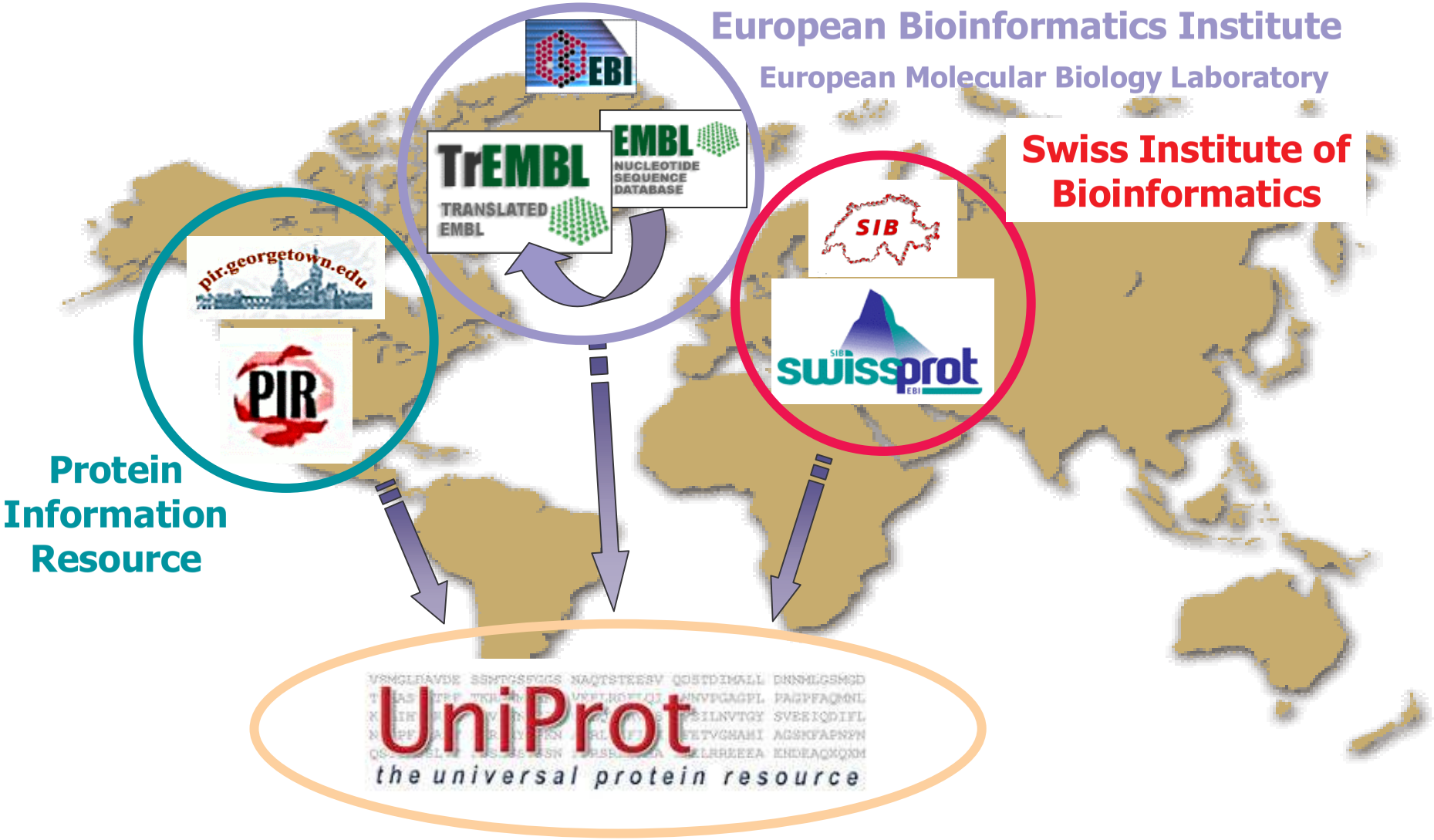
Register now! Deadline June 30, 2006

REGISTRATION
POSTERS
PROGRAM
SPEAKERS
SPONSORS & EXHIBITORS
ACCOMMODATION & TRAVEL
VENUE
ISMB 2006
LINKS
CONTACT

Confirmed Speakers:

Ron Appel (CHE)	Michael Dunn (USA)	Rich Roberts (USA)
Rolf Apweiler (GBR)	Ernest Feytmans (CHE)	Burkhard Rost (USA)
Michael Ashburner (USA)	Takashi Gojobori (JAP)	Kenneth Rudd (USA)
Terri Attwood (GBR)	Michael Gribskov (USA)	Chris Sander (USA)
Amos Bairoch (CHE)	Winston Hide (ZAF)	Torsten Schwede (CHE)
Alex Bateman (GBR)	Des Higgins (IRL)	Joel Sussman (ISR)
Ewan Birney (GBR)	Denis Hochstrasser (CHE)	Janet Thornton (GBR)
Judith Blake (USA)	Victor Jongeneel (CHE)	Jean-François Tomb (USA)
Peer Bork (DEU)	Daniel Kahn (FRA)	Vitek Tracz
Philip Bourne (USA)	Minoru Kanehisa (JAP)	Gunnar von Heijne (SWE)
Steven Brenner (USA)	Jack Leunissen (NLD)	Owen White (USA)
Søren Brunak (DEN)	Kenta Nakai (JAP)	Edgar Wingender (GER)
Doug Brutlag (USA)	Cedric Notredame (FRA)	Shoshana Wodak (CAN)
Philipp Bucher (CHE)	Christine Orengo (GBR)	Cathy Wu (USA)
Jean-Michel Claverie (FRA)	Christos Ouzounis (GBR)	Ioannis Xenarios (CHE)
Julio Collado-Vides (MEX)	William Pearson (USA)	
Antoine Danchin (FRA)	Manuel Peitsch (CHE)	

The UniProt consortium



The UniProt Consortium

UniProt (Universal Protein Resource): the world's most comprehensive catalog of information on proteins

www.uniprot.org, Wu et al. Nucleic Acids Res. 34:D187-191(2006).

UniProt Knowledgebase
UniRef clusters (100/90/50% identity)
UniParc (UniProt Archive)

UniProt: What does it mean for Swiss-Prot users?

- Swiss-Prot and TrEMBL will continue to exist under those names (although now called UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- PIR has stopped protein sequence db (PIR-PSD) activity but concentrates on other topics (UniRef, IProClass)
- Information unique to PIR-PSD has been integrated into Swiss-Prot/TrEMBL

=> UniProt Knowledgebase: Swiss-Prot and TrEMBL

- Biweekly releases

The UniProt groups from SIB, EBI and PIR (Antibes, September 2004)

In Geneva (SIB):

2 Group Leaders
42 Annotators
4 Prosite annotators
18 Programmers and Researchers
5 Administrators, science communicators
3 System Administrators
4 Students
---
78 people

At EBI:

(Swiss-Prot + EMBL + TrEMBL)
75 people (29 Annotators)

At PIR:

1 Group Leader
13 Protein Science Team
12 Informatics Team
---
26 people

UniProt databases

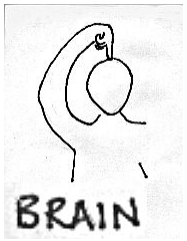
UniProt KnowledgeBase

swissprot

TrEMBL TRANSLATED EMBL

UniProtKB/TrEMBL
Computer annotated protein sequences

Release 33.1 of 13-Jun-2006: 2965756 entries



UniProtKB/Swiss-Prot
Manually annotated protein sequences

Release 50.1 of 13-Jun-2006: 223100 entries

UniRef

UniRef100
UniRef 90
UniRef 50

UniRef
Non-redundant Reference
100%>90%>50%

- One **UniRef100** entry = **All identical sequences** (including fragments).
- One **UniRef90** entry = Sequences that have at least **90% or more identity**.
- One **UniRef50** entry = Sequences that are at least **50% identical**.

Independently of the species.

UniParc

UniParc
The UniProt Archive

UniProt Archives: Archived raw protein sequences, found in publicly accessible databases:

Swiss-Prot, TrEMBL, PIR, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, Patent Offices.

Use with extreme caution: Contains pseudogenes, incorrect CDS predictions, etc...

UniProtKB Release 8.1 consists of:

UniProt Knowledgebase

- Swiss-Prot: Manually annotated section
- TrEMBL: Automatically generated section

UniProtKB/Swiss-Prot

- Swiss-Prot created in July 1986;
- Collaboration of the SIB and the EMBL/EBI;
- Annotated, non-redundant, cross-referenced, documented protein sequence and knowledge database;
- Release **50.1** of Swiss-Prot contains **223'100** sequence entries, abstracted from 143'000 references. This represents 240'000 sequences, taking into account splice variants; 2'700'000 cross-references; flat file size 900 MegaBytes;
- Biweekly releases; available from about ~100 servers, the main sources being ExPASy and www.uniprot.org

The contents of the Swiss-Prot protein sequence database

- Sequences!
- **ANNOTATIONS**
- References
- Taxonomic d
- Keywords
- Cross-referer
- Documentati

- Function(s); role(s)
- Post-translational modifications
- Domains
- Subcellular location
- Protein/protein interactions
- Similarities
- Diseases, mutagenesis
- Conflicts and variants

In-Silico Analysis of Proteins,
Celebrating the 20th Anniversary of Swiss-Prot
Register now!

Swiss-Prot
Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:

- **UniProtKB/Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].
- **UniProtKB/TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Knowledgebase Release 8.1 consists of:
UniProtKB/Swiss-Prot Release 50.1 of 13-Jun-2006: 223100 entries (More statistics)
UniProtKB/TrEMBL Release 33.1 of 13-Jun-2006: 2965756 entries (More statistics)

> Swiss-Prot headlines
Man gave names to all the... proteins (Read more...)

Access to the UniProt Knowledgebase

- **SRS** - Access to UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and other databases using the Sequence Retrieval System
- Full text search in the UniProt Knowledgebase
- Advanced search in the UniProt Knowledgebase by description, gene name and organism (can be used to create html links to UniProt Knowledgebase queries)
- Taxonomy browser (NEWT)
- **BLAST** similarity search

- by description or identification (any word in the DE, OS, OG, GN and ID lines)
- by citation (RL line; UniProtKB/Swiss-Prot only)

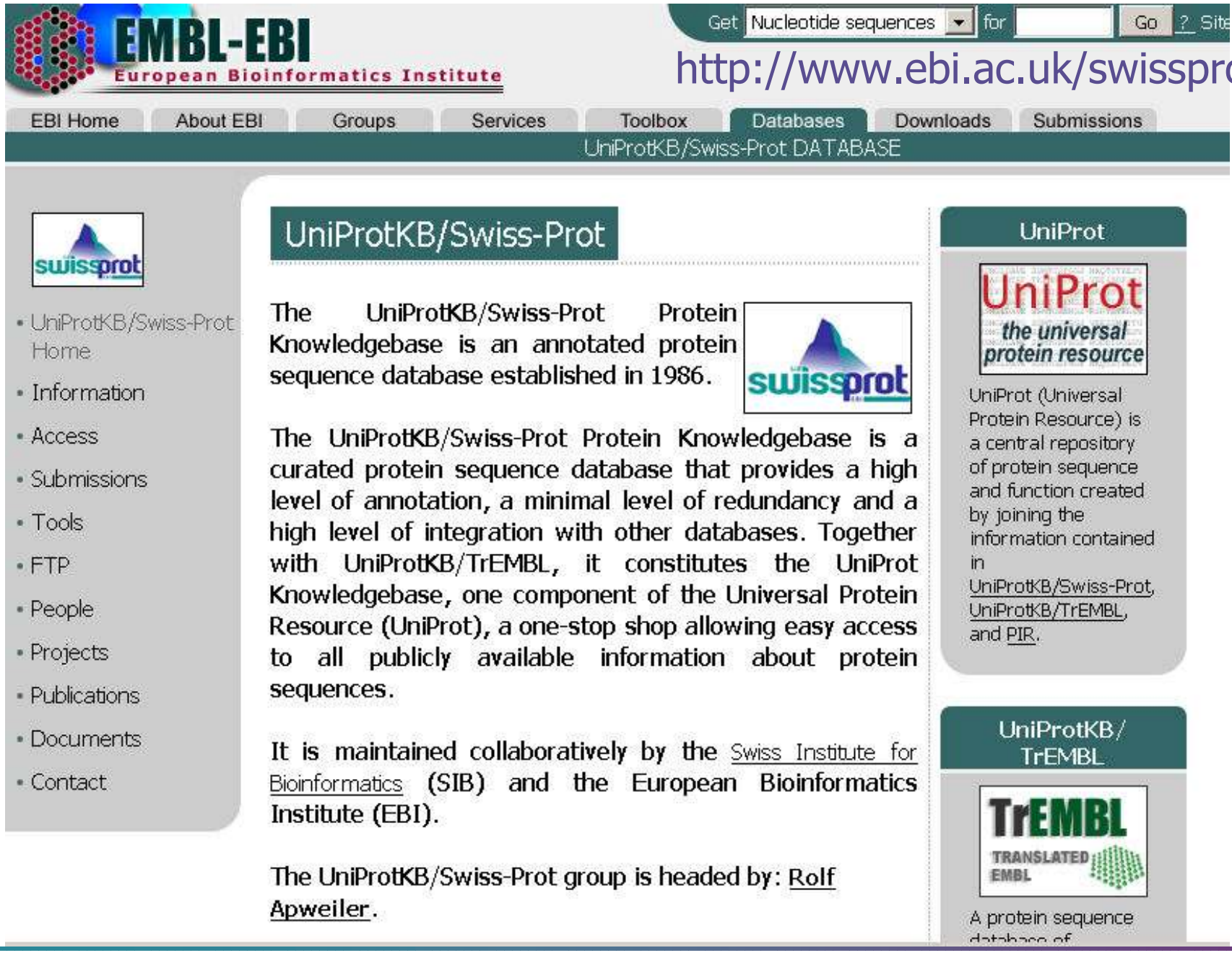
- Retrieve a list of UniProtKB entries
- Randomly retrieve a UniProtKB entry

- UniProtKB Sequence/Annotation Version Database new
- Swiss-Prot ID tracker

http://www.expasy.org/sprot/

Documents and services

- Swiss-Prot documents - user manual, release notes, indices and lots of other **important** documents and lists


EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site
http://www.ebi.ac.uk/swissprot/
EBI Home
About EBI
Groups
Services
Toolbox
Databases
Downloads
Submissions
UniProtKB/Swiss-Prot DATABASE
swissprot
UniProtKB/Swiss-Prot Home
Information
Access
Submissions
Tools
FTP
People
Projects
Publications
Documents
Contact
UniProtKB/Swiss-Prot
The UniProtKB/Swiss-Prot Protein Knowledgebase is an annotated protein sequence database established in 1986.
swissprot
The UniProtKB/Swiss-Prot Protein Knowledgebase is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and a high level of integration with other databases. Together with UniProtKB/TrEMBL, it constitutes the UniProt Knowledgebase, one component of the Universal Protein Resource (UniProt), a one-stop shop allowing easy access to all publicly available information about protein sequences.
It is maintained collaboratively by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI).
The UniProtKB/Swiss-Prot group is headed by: Rolf Apweiler.
UniProt
UniProt
the universal protein resource
UniProt (Universal Protein Resource) is a central repository of protein sequence and function created by joining the information contained in UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, and PIR.
UniProtKB/ TrEMBL
TrEMBL
TRANSLATED EMBL
A protein sequence

UniProt
the universal protein resource

Text Search UniProt Knowledgebase

Home | About UniProt | Getting Started | Searches/Tools | Databases | Support/Documentation

www.uniprot.org

Welcome to UniProt

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

UniProt is comprised of three components, each optimized for different uses. The **UniProt Knowledgebase (UniProt)** is the central access point for extensive curated protein information, including function, classification, and cross-reference. The **UniProt Non-redundant Reference (UniRef)** databases combine closely related sequences into a single record to speed searches. The **UniProt Archive (UniParc)** is a comprehensive repository, reflecting the history of all protein sequences.

The sequences and information in UniProt are accessible via **text search**, **BLAST similarity search**, and **FTP**.

European Bioinformatics Institute

Swiss Institute of Bioinformatics

Georgetown University

swissprot

Let's now look at the content of a Swiss-Prot entry

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

Manual annotation

```
ID   SYI_ARCFU      STANDARD;      PRT;  1018 AA.
AC   O29622;
DT   15-JUL-1998, integrated into UniProtKB/Swiss-Prot.
DT   01-JAN-1998, sequence version 1.
DT   27-JUN-2006, entry version 43.
DE   Isoleucyl-tRNA synthetase (EC 6.1.1.5) (Isoleucine--tRNA ligase)
DE   (IleRS).
GN   Name=ileS; OrderedLocusNames=AF_0633;
OS   Archaeoglobus fulgidus.
OC   Archaea; Euryarchaeota; Archaeoglobi; Archaeoglobales;
OC   Archaeoglobaceae; Archaeoglobus.
OX   NCBI_TaxID=2234;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RC   STRAIN=VC-16 / DSM 4304 / ATCC 49558 / JCM 9628;
RX   MEDLINE=98049343; PubMed=9389475; DOI=10.1038/37052;
RA   Klenk H.-P., Clayton R.A., Tomb J.-F., White O., Nelson K.E.,
RA   Ketchum K.A., Dodson R.J., Gwinn M.L., Hickey E.K., Peterson J.D.,
RA   Richardson D.L., Kerlavage A.R., Graham D.E., Kyrpides N.C.,
RA   Fleischmann R.D., Quackenbush J., Lee N.H., Sutton G.G., Gill S.R.,
RA   Kirkness E.F., Dougherty B.A., McKenney K., Adams M.D., Loftus B.J.,
RA   Peterson S.N., Reich C.I., McNeil L.K., Badger J.H., Glodek A.,
RA   Zhou L., Overbeek R., Gocayne J.D., Weidman J.F., McDonald L.A.,
RA   Utterback T.R., Cotton M.D., Spriggs T., Artiach P., Kaine B.P.,
RA   Sykes S.M., Sadow P.W., D'Andrea K.P., Bowman C., Fujii C.,
RA   Garland S.A., Mason T.M., Olsen G.J., Fraser C.M., Smith H.O.,
RA   Woese C.R., Venter J.C.;
RT   "The complete genome sequence of the hyperthermophilic, sulphate-
RT   reducing archaeon Archaeoglobus fulgidus.";
RL   Nature 390:364-370(1997).
CC   -!- FUNCTION: Catalyzes the attachment of isoleucine to tRNA(Ile). As
CC       IleRS can inadvertently accommodate and process structurally
CC       similar amino acids such as valine, to avoid such errors it has
CC       two additional distinct tRNA(Ile)-dependent editing activities.
CC       One activity is designated as 'pretransfer' editing and involves
CC       the hydrolysis of activated Val-AMP. The other activity is
CC       designated 'posttransfer' editing and involves deacylation of
CC       mischarged Val-tRNA(Ile) (By similarity).
CC   -!- CATALYTIC ACTIVITY: ATP + L-isoleucine + tRNA(Ile) = AMP +
CC       diphosphate + L-isoleucyl-tRNA(Ile).
CC   -!- COFACTOR: Zinc (By similarity).
CC   -!- SUBUNIT: Monomer (By similarity).
CC   -!- SUBCELLULAR LOCATION: Cytoplasm (By similarity).
CC   -!- DOMAIN: IleRS has two distinct active sites: one for
CC       aminoacylation and one for editing. The misactivated valine is
```

Original text format, as described in the Swiss-Prot user manual http://www.expasy.org/sprot/userman.html, or as available by ftp, or from many web servers...

.... But there is a more user-friendly way of looking at all this data!

NiceProt

- a user-friendly, added-value view of a UniProtKB entry on ExPASy
- Many additional html links to related servers, databases and documentation
- Direct submission to analysis tools, minimising the number of clicks and typing
- The default view for a Swiss-Prot entry on ExPASy

Accession number: to be used if you have to cite a Swiss-Prot entry in your publication (never cite the entry name (ID) alone)

UniProtKB/Swiss-Prot entry P00740

Direct (fast) BLASTP submission

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

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

Entry information	
Entry name	FA9_HUMAN
Primary accession number	P00740
Secondary accession numbers	None
Integrated into Swiss-Prot on	July 21, 1986
Sequence was last modified on	June 7, 2005 (Sequence version 2)
Annotations were last modified on	June 13, 2006 (Entry version 110)
Name and origin of the protein	
Protein name	**Coagulation factor IX [Precursor]**
Synonyms	**EC 3.4.21.22** **Christmas factor** **Plasma thromboplastin component** **PTC**
Contains	**Coagulation factor IXa light chain** **Coagulation factor IXa heavy chain**
Gene name	**Name: F9**
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

References

[1] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=86000558; PubMed=2994716; [NCBI, ExPASy, EBI, Israel, Japan]
Yoshitake S., Schach B.G., Foster D.C., Davie E.W., Kurachi K.;
"Nucleotide sequence of the gene for human factor IX (antihemophilic factor B).";
Biochemistry 24:3736-3750(1985).

Sequence (DNA or protein)

[2] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=85190593; PubMed=3857619; [NCBI, ExPASy, EBI, Israel, Japan]
McGraw R.A., Davis L.M., Noyes C.M., Lundblad R.L., Roberts H.R., Graham J.B., Stafford D.W.;
"Evidence for a prevalent dimorphism in the activation peptide of human coagulation factor IX.";
Proc. Natl. Acad. Sci. U.S.A. 82:2847-2851(1985).

[3] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=84236100; PubMed=6329734; [NCBI, ExPASy, EBI, Israel, Japan]
Anson D.S., Choo K.H., Rees D.J.G., Giannelli F., Gould K.G., Huddleston J.A., Brownlee G.G.;
"The gene structure of human anti-haemophilic factor IX.";
EMBO J. 3:1053-1060(1984).

[4] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=83220788; PubMed=6687940; [NCBI, ExPASy, EBI, Israel, Japan]
Jaye M., de la Salle H., Schamber F., Balland A., Kohli V., Findeli A., Tolstoshev P., Lecocq J.P.;
"Isolation of a human anti-haemophilic factor IX cDNA clone using a unique 52-base synthetic oligonucleotide probe deduced from the amino acid sequence of bovine factor IX.";
Nucleic Acids Res. 11:2325-2335(1983).

Sequence variants

[5] SEQUENCE FROM NUCLEIC ACID AND VARIANTS THR-194 AND PRO-461.
Rieder M.J., Armel T.Z., Carrington D.P., Ozuna M., Kuldanek S.A., Rajkumar N.R., Toth E.J., Yi Q., Nickerson D.A.;
Submitted (AUG-2002) to the EMBL/GenBank/DDBJ databases.

...12 references omitted...

Post-translational modifications

[17] STRUCTURE OF CARBOHYDRATE ON SER-107.
MEDLINE=92388094; PubMed=1517205; [NCBI, ExPASy, EBI, Israel, Japan]
Nishimura H., Takao T., Hase S., Shimonishi Y., Iwanaga S.;
"Human factor IX has a tetrasaccharide O-glycosidically linked to serine 61 through the fucose residue.";
J. Biol. Chem. 267:17520-17525(1992).

[18] PHOSPHORYLATION OF SER-114.
Harris R.J., Papac D.I., Truong L., Smith K.J.;
"Partial phosphorylation of serine-68 in EGF-1 of human factor IX.";
(In) Abstracts of XIth international conference on methods in protein structure analysis, pp.50-50, Annecy (1996).

[19] POST-TRANSLATIONAL MODIFICATIONS.
MEDLINE=20575397; PubMed=11133752; [NCBI, ExPASy, EBI, Israel, Japan]
Arruda V.R., Hagstrom J.N., Deitch J., Heiman-Patterson T., Camire R.M., Chu K., Fields P.A., Herzog R.W., Couto L.B., Larson P.J., High K.A.;
"Posttranslational modifications of recombinant myotube-synthesized human factor IX.";
Blood 97:130-138(2001).

3D structure

[20] STRUCTURE BY NMR OF 47-93.
MEDLINE=95229607; PubMed=7713897; [NCBI, ExPASy, EBI, Israel, Japan]
Freedman S.J., Furie B.C., Furie B., Baleja J.D.;

Comments: « structured free text », 27 defined topics

Comments

- ***FUNCTION***: Factor IX is a vitamin K-dependent plasma protein that participates in the intrinsic pathway of blood coagulation by converting factor X to its active form in the presence of Ca(2+) ions, phospholipids, and factor VIIIa.
- ***CATALYTIC ACTIVITY***: Selective cleavage of Arg-|-Ile bond in factor X to form factor Xa.
- ***SUBUNIT***: Heterodimer of a light chain and a heavy chain; disulfide-linked.
- ***SUBCELLULAR LOCATION***: Secreted protein.
- ***TISSUE SPECIFICITY***: Synthesized primarily in the liver and secreted in plasma.
- ***DOMAIN***: Calcium binds to the gamma-carboxyglutamic acid (Gla) residues and, with stronger affinity, to another site, beyond the Gla domain.
- ***PTM***: Activated by factor XIa, which excises the activation peptide.
- ***DISEASE***: Defects in F9 are the cause of recessive X-linked hemophilia B (HEMB) [MIM:306900]; also known as Christmas disease.
- ***DISEASE***: Mutations in position 43 (Oxford-3, San Dimas) and 46 (Cambridge) prevents cleavage of the propeptide, mutation in position 93 (Alabama) probably fails to bind to cell membranes, mutation in position 191 (Chapel-Hill) or in position 226 (Nagoya OR Hilo) prevent cleavage of the activation peptide.
- ***PHARMACEUTICAL***: Available under the names BeneFix (Baxter and American Home Products). Used to treat hemophilia B.
- ***MISCELLANEOUS***: In 1952, one of the earliest researchers of the disease, Dr. R.G. Macfarlane used the patient's surname, Christmas, to refer to the disease and also to refer to the clotting factor which he called the 'Christmas Factor' At the time Stephen Christmas was a 5-year-old boy. He died in 1993 at the age of 46 from acquired immunodeficiency syndrome contracted through treatment with blood products.
- ***SIMILARITY***: Belongs to the peptidase S1 family [vi
- ***SIMILARITY***: Contains 2 EGF-like domains.
- ***SIMILARITY***: Contains 1 Gla (gamma-carboxy-glut
- ***SIMILARITY***: Contains 1 peptidase S1 domain [view
- ***WEB RESOURCE***: NAME=HAEMB; NOTE=Hemophili URL="http://www.kcl.ac.uk/ip/petergreen/haemBdat
- ***WEB RESOURCE***: NAME=BeneFix; NOTE=Clinical in URL="http://www.wyeth.com/products/benefix.asp"
- ***WEB RESOURCE***: NAME=Protein Spotlight; NOTE= URL="http://www.expasy.org/spotlight/back_issues/
- ***WEB RESOURCE***: NAME=GeneReviews; URL="http

Manually annotated

Information from papers,
specialized databases, computer prediction,
external experts, brain storming

Distinction between data obtained
experimentally and computerized inferences

Cross-references

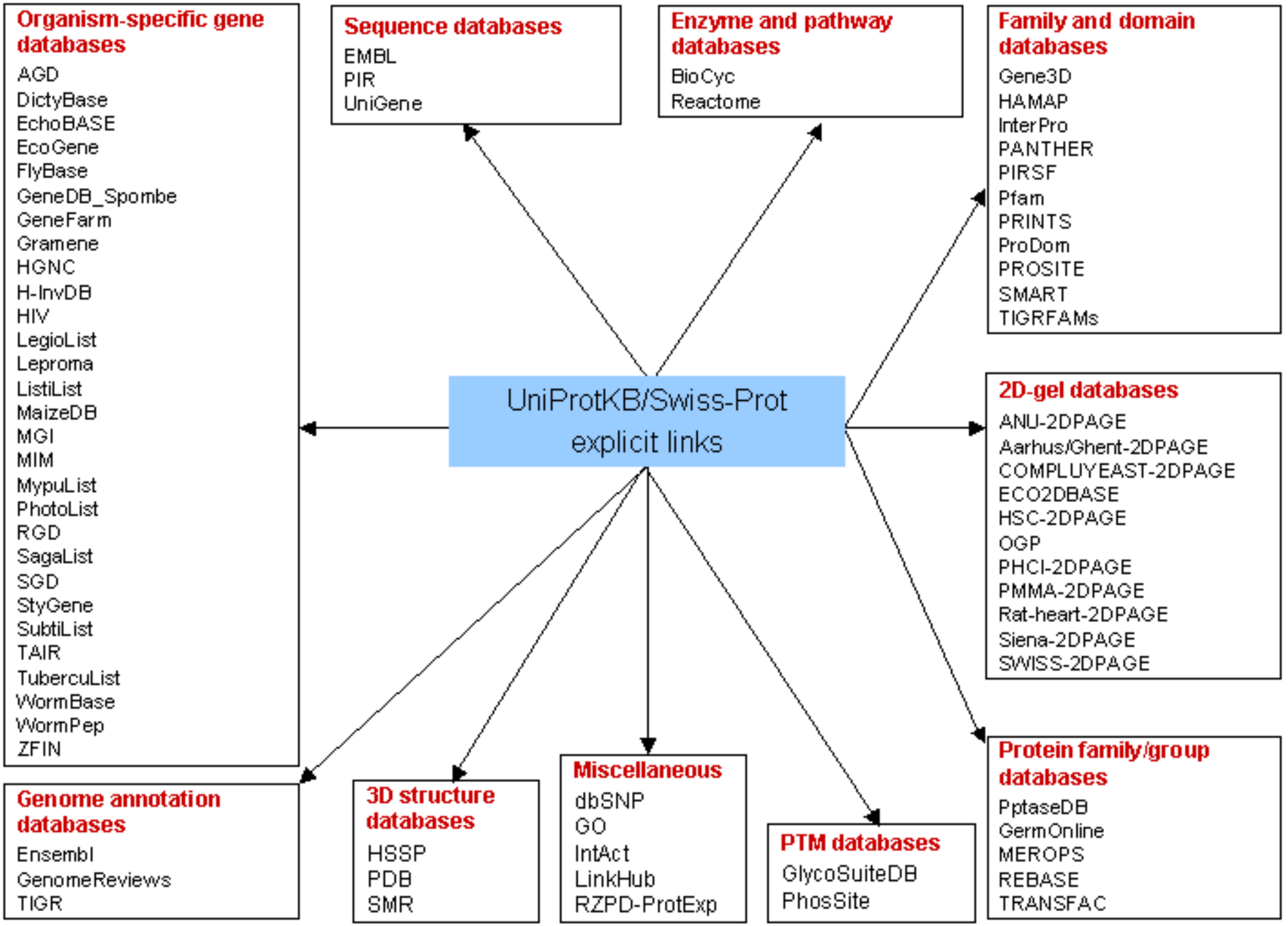
Gasteiger E. et al, Curr. Issues Mol. Biol. 3:47-55(2001)
www.expasy.org/cgi-bin/lists?dbxref.txt

- Swiss-Prot was the first database with X-ref.;
- Explicitly X-referenced to 74 databases:
  - DNA (EMBL/GenBank/DDBJ),
  - 3D-structure (PDB)
  - Family and domain (InterPro, PROSITE, Pfam, Prints, etc.)
  - genomic (MIM, MGI, FlyBase, SGD, SubtiList, etc.)
  - 2D-gel (e.g. SWISS-2DPAGE)
  - specialized db (e.g.GlycoSuiteDB, PhosSite, MEROPS);
  - literature (PubMed)
- Each UniProtKB/Swiss-Prot entry can be seen as a central hub for the data available about the protein it describes

Implicit Cross-references on ExPASy

Implicit X-references to 27 additional db added by the ExPASy server on the WWW (i.e.: GeneCards, ModBase, etc.)

These X-refs are *not present as hard-coded DR lines* in the Swiss-Prot entry as it can be downloaded by ftp, but are *added on the fly* when someone views an entry on ExPASy. This can be done because enough information is present in the Swiss-Prot entry to access the related information in another db. Example: All Swiss-Prot/TrEMBL are linked to the BLOCKS domain db, via the Swiss-Prot/TrEMBL accession number


Organism-specific gene databases
AGD
DictyBase
EchoBASE
EcoGene
FlyBase
GeneDB_Spombe
GeneFarm
Gramene
HGNC
H-InvDB
HIV
LegioList
Leproma
ListiList
MaizeDB
MGI
MIM
MypuList
PhotoList
RGD
SagaList
SGD
StyGene
SubtiList
TAIR
TubercuList
WormBase
WormPep
ZFIN
Sequence databases
EMBL
PIR
UniGene
Enzyme and pathway databases
BioCyc
Reactome
Family and domain databases
Gene3D
HAMAP
InterPro
PANTHER
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs
UniProtKB/Swiss-Prot explicit links
2D-gel databases
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
ECO2DBASE
HSC-2DPAGE
OGP
PHCI-2DPAGE
PMMA-2DPAGE
Rat-heart-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE
Genome annotation databases
Ensembl
GenomeReviews
TIGR
3D structure databases
HSSP
PDB
SMR
Miscellaneous
dbSNP
GO
IntAct
LinkHub
RZPD-ProtExp
PTM databases
GlycoSuiteDB
PhosSite
Protein family/group databases
PptaseDB
GermOnline
MEROPS
REBASE
TRANSFAC

Cross-references		
Sequence databases		
EMBL	K02402; AAB59620.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	J00136; AAA98726.1; -; mRNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	J00137; AAA52763.1; ALT_SEQ; mRNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	M11309; AAA52023.1; ALT_SEQ; mRNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	K02053; AAA56822.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	K02048; AAA56822.1; JOINED; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	K02049; AAA56822.1; JOINED; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	K02051; AAA56822.1; JOINED; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	K02052; AAA56822.1; JOINED; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	AF536327; AAM96188.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	M35672; AAA51981.1; -; mRNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	M19063; AAA52456.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	S66752; AAB28588.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
	S68634; AAB29758.1; -; Genomic_DNA.	[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A00922; KFHU.	
UniGene	Hs.522798	
3D structure databases		
PDB	1CFH; NMR; @=47-93.	[ExPASy / RCSB / EBI]
	1CFI; NMR; @=47-93.	[ExPASy / RCSB / EBI]
	1EDM; X-ray; B/C=92-130.	[ExPASy / RCSB / EBI]
	1IXA; NMR; @=92-130.	[ExPASy / RCSB / EBI]
	1MGX; NMR; @=47-93.	[ExPASy / RCSB / EBI]
	1NL0; X-ray; G=47-91.	[ExPASy / RCSB / EBI]
	1RFN; X-ray; A=227-461, B=133-188.	[ExPASy / RCSB / EBI]
	Detailed list of linked structures.	
SMR	P00740; 47-191.	
ModBase	P00740.	
Protein-protein interaction databases		
DIP	P00740.	
Protein family/group databases		
MEROPS	S01.214; -.	

PTM databases	
GlycoSuiteDB	P00740; -.
Enzyme and pathway databases	
Reactome	P00740; -.
Polymorphism databases	
SeattleSNPs	F9.
2D gel databases	
SWISS-2DPAGE	Get region on 2D PAGE.
Organism-specific gene databases	
HGNC	HGNC:3551; F9.
GeneCards	F9.
GeneLynx	F9; Homo sapiens.
GenAtlas	F9.
MIM	306900; gene+phenotype. [NCBI / EBI]
HOVERGEN	[Family / Alignment / Tree]
Gene expression databases	
CleanEx	HGNC:3551; F9.
Ontologies	
GO	GO:0005576; Cellular component: extracellular region *(non-traceable author statement).* GO:0003803; Molecular function: coagulation factor IXa activity *(traceable author statement).* GO:0007596; Biological process: blood coagulation *(traceable author statement).* QuickGo view.
Family and domain databases	
InterPro	IPR000152; Asx_hydroxyl_S. IPR002383; Coagulation_factor_Gla. IPR006210; EGF. IPR001438; EGF_2. IPR000742; EGF_3. IPR001881; EGF_Ca_bd. IPR006209; EGF_like. IPR013032; EGF_like_reg. IPR012224; Pept_S1A_FX. IPR009003; Pept_Ser_Cys.

Pfam	PF00594; Gla; 1. PF00089; Trypsin; 1. Pfam graphical view of domain structure.
PIRSF	PIRSF001143; Factor_X; 1.
PRINTS	PR00722; CHYMOTRYPSIN. PR00010; EGFBLOOD. PR00001; GLABLOOD.
SMART	SM00181; EGF; 2. SM00179; EGF_CA; 1. SM00069; GLA; 1. SM00020; Tryp_SPc; 1. SMART graphical view of domain structure.
PROSITE	PS00010; ASX_HYDROXYL; 1. PS00022; EGF_1; 1. PS01186; EGF_2; 2. PS50026; EGF_3; 1. PS01187; EGF_CA; 1. PS00011; GLA_1; 1. PS50998; GLA_2; 1. PS50240; TRYPSIN_DOM; 1. PS00134; TRYPSIN_HIS; 1. PS00135; TRYPSIN_SER; 1. PROSITE graphical view of domain structure (profiles).
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
BLOCKS	P00740.
Genome annotation databases	
Ensembl	ENSG00000101981; Homo sapiens. [Contig view]
Other	
LinkHub	P00740; -.
RZPD-ProtExp	IOH41673; -. T0161; -.
SOURCE	F9; Homo sapiens.
ProtoNet	P00740.
UniRef	View cluster of proteins with at least 50% / 90% / 100% identity.

Key	From	To	Length	Description
SIGNAL	1	28	28	*Potential.*
PROPEP	29	46	18	
CHAIN	47	461	415	Coagulation factor IX.
CHAIN	47	191	145	Coagulation factor IXa light chain.
PROPEP	192	226	35	Activation peptide.
CHAIN	227	461	235	Coagulation factor IXa heavy chain.
DOMAIN	47	92	46	Gla.
DOMAIN	93	129	37	EGF-like 1, calcium-binding *(Potential)*.
DOMAIN	130	171	42	EGF-like 2.
DOMAIN	227	461	235	Serine protease.
SITE	191	192	2	Cleavage (by factor XIa).
SITE	226	227	2	Cleavage (by factor XIa).
MOD_RES	53	53		4-carboxyglu
MOD_RES	54	54		4-carboxyglu
MOD_RES	61	61		4-carboxyglu
MOD_RES	63	63		4-carboxyglu
MOD_RES	66	66		4-carboxyglu
MOD_RES	67	67		4-carboxyglu
MOD_RES	72	72		4-carboxyglu
MOD_RES	73	73		4-carboxyglu
MOD_RES	76	76		4-carboxyglu
MOD_RES	79	79		4-carboxyglu
MOD_RES	82	82		4-carboxyglu
MOD_RES	86	86		4-carboxyglu
MOD_RES	110	110		3-hydroxyasp
MOD_RES	114	114		Phosphoseri
MOD_RES	201	201		Sulfotyrosi
MOD_RES	204	204		Phosphoseri
DISULFID	64	69		
DISULFID	97	108		
DISULFID	102	117		
DISULFID	119	128		
DISULFID	134	145		*By similari*

VARIANT	419	419	1	V -> E (in HEMB; moderately severe).	VAR_006606
VARIANT	424	424	1	F -> V (in HEMB).	VAR_006607
VARIANT	426	426	1	T -> P (in HEMB; severe; Barcelos).	VAR_006608
VARIANT	430	430	1	S -> T (in HEMB).	VAR_006609
VARIANT	431	431	1	W -> G (in HEMB).	VAR_006610
VARIANT	431	431	1	W -> R (in HEMB; moderate).	VAR_006611
VARIANT	432	432	1	G -> S (in HEMB; severe).	VAR_006612
VARIANT	432	432	1	G -> V (in HEMB; severe).	VAR_006613
VARIANT	433	433	1	E -> A (in HEMB).	VAR_006614
VARIANT	433	433	1	E -> K (in HEMB).	VAR_006615
VARIANT	435	435	1	C -> Y (in HEMB).	VAR_017364
VARIANT	436	436	1	A -> V (in HEMB; moderately severe; Niigata).	VAR_006616
VARIANT	442	442	1	G -> E (in HEMB).	VAR_017365
VARIANT	442	442	1	G -> R (in HEMB; severe; Angers).	VAR_017322
VARIANT	443	443	1	I -> T (in HEMB; moderately severe; Long Beach, Los Angeles and Vancouver).	VAR_017323
VARIANT	445	445	1	T -> TIYT (in HEMB; severe; Lousada).	VAR_006617
VARIANT	449	449	1	R -> Q (in HEMB; mild).	VAR_006618
VARIANT	449	449	1	R -> W (in HEMB; mild).	VAR_006619
VARIANT	450	450	1	Y -> C (in HEMB; severe).	VAR_006620
VARIANT	453	453	1	W -> R (in HEMB).	VAR_017324
VARIANT	454	454	1	I -> T (in HEMB; Italy).	VAR_006621
VARIANT	461	461	1	T -> P.	VAR_014308
HELIX	60	63	4		
TURN	64	65	2		
HELIX	71	77	7		
TURN	78	78	1		
HELIX	81	88	8		
TURN	89	91	3		
TURN	134	135	2		

Feature aligner

Feature table viewer

Keywords

3D-structure; Blood coagulation; Calcium; Calcium-binding; Direct protein sequencing; Disease mutation; EGF-like domain; Gamma-carboxyglutamic acid; Glycoprotein; Hemophilia; Hydrolase; Hydroxylation; Pharmaceutical; Phosphorylation; Plasma; Polymorphism; Protease; Repeat; Serine protease; Signal; Sulfation; Zymogen.

UniProt Knowledgebase keyword: Blood coagulation

Description

Protein involved in blood clotting, a complex enzymatic cascade, in which the activated form of one factor catalyzes the activation of the next factor. Both, the extrinsic clotting pathway, induced by a damaged surface, and the intrinsic pathway, induced by a trauma, converge in a final common pathway to form cross-linked fibrin clots.

Gene ontology links

GO:0007596; blood coagulation.

Hierarchy

Biological process ⊃ Blood coagulation

More specific sets of entries

Fibrinolysis, Hemophilia, Hemostasis, Thrombophilia, von W

Keyword definition and usage in Swiss-Prot

Linked to Gene Ontology to further facilitate information retrieval via controlled vocabularies

Entries in UniProtKB/Swiss-Prot (352):

Send selected sequences to Retrieve sequences (FASTA format) Submit Query Select all

	Entry name	AC	Gene names	Description	Organisms	Length
☐	ACH1_LONAC	P23604		**Achelase-1** (EC 3.4.21.-) (Achelase I)	Lonomia achelous (Giant silkworm moth) (Saturnid moth)	213
☐	ACH2_LONAC	P23605		**Achelase-2** (EC 3.4.21.-) (Achelase II)	Lonomia achelous (Giant silkworm moth) (Saturnid moth)	214
☐	ANT3_BOVIN	P41361	**SERPINC1**, AT3	**Antithrombin-III** (ATIII)	Bos taurus (Bovine)	433
☐	ANT3_CHICK	Q03352	**SERPINC1**, AT3	**Antithrombin-III precursor** (ATIII) (Fragment)	Gallus gallus (Chicken)	105
☐	ANT3_HUMAN	P01008	**SERPINC1**, AT3, *PRO0309*	**Antithrombin-III precursor** (ATIII)	Homo sapiens (Human)	464
☐	ANT3_MESAU	P81050	**SERPINC1**, AT3	**Antithrombin-III** (ATIII) (Fragment)	Mesocricetus auratus (Golden hamster)	25

Sequence information

Length: **461 AA** [This is the length of the unprocessed precursor]	Molecular weight: **51778 Da** [This is the MW of the unprocessed precursor]	CRC64: **C4720C1234477EF5** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MQRVNMIMAE SPGLITICLL GYLLSAECTV FLDHENANKI LNRPKRYNSG KLEEFVQGNL

        70         80         90        100        110        120
ERECMEEKCS FEEAREVFEN TERTTEFWKQ YVDGDQCESN PCLNGGSCKD DINSYECWCP

       130        140        150        160        170        180
FGFEGKNCEL DVTCNIKNGR CEQFCKNSAD NKVVCSCTEG YRLAENQKSC EPAVPFPCGR

       190        200        210        220        230        240
VSVSQTSKLT RAETVFPDVD YVNSTEAETI LDNITQSTQS FNDFTRVVGG EDAKPGQFPW

       250        260        270        280        290        300
QVVLNGKVDA FCGGSIVNEK WIVTAAHCVE TGVKITVVAG EHNIEETEHT EQKRNVIRII

       310        320        330        340        350        360
PHHNYNAAIN KYNHDIALLE LDEPLVLNSY VTPICIADKE YTNIFLKFGS GYVSGWGRVF

       370        380        390        400        410        420
HKGRSALVLQ YLRVPLVDRA TCLRSTKFTI YNNMFCAGFH EGGRDSCQGD SGGPHVTEVE

       430        440        450        460
GTSFLTGIIS WGEECAMKGK YGIYTKVSRY VNWIKEKTKL T
```

P00740 in FASTA format

… and many useful links:

View entry in original UniProtKB/Swiss-Prot format
View entry in raw text format (no links)
Report form for errors/updates in this UniProtKB/Swiss-Prot entry

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

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

ScanProsite, MotifScan

Submit a homology modeling request to SWISS-MODEL

NPSA Sequence analysis tools

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

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

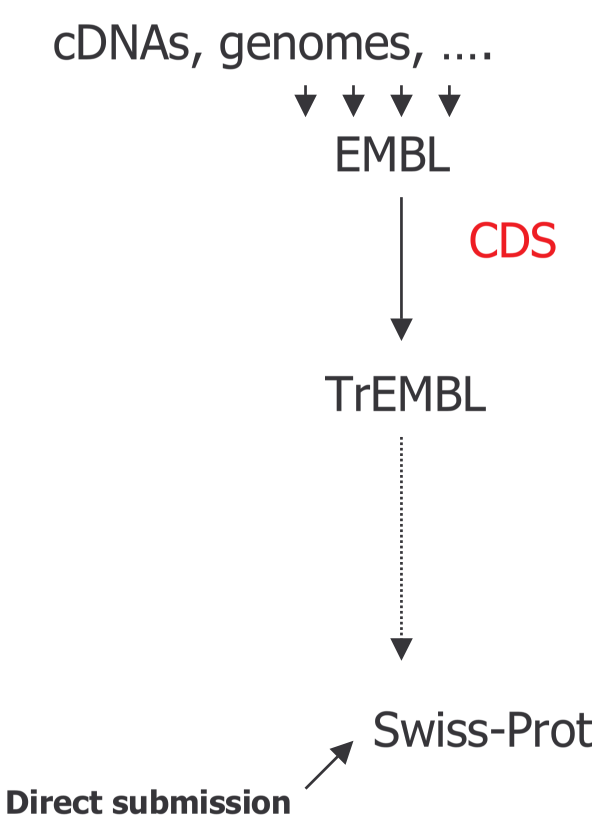
An example of a protein with many different names (ANXA5_HUMAN; P08758)

DE Annexin A5 (Annexin V) (Lipocortin V) (Endonexin II) (Calphobindin I)
DE (CBP-I) (Placental anticoagulant protein I) (PAP-I) (PP4)
DE (Thromboplastin inhibitor) (Vascular anticoagulant-alpha) (VAC-alpha)
DE (Anchorin CII).

An example of a protein with many different gene names (TUP1_YEAST; P16649)

GN Name=TUP1;
GN Synonyms=AAR1, AER2, AMM1, CYC9, FLK1, SFL2, UMR7;
GN OrderedLocusNames=YCR084C;
GN ORFNames=YCR84C;

The simplified story of a Swiss-Prot entry



« **Automatic** »
- Redundancy check (100% merge)
- Family attribution (InterPro)
- Many other cross-references
- Annotation (computer)

« **Manual** »
- Redundancy (merge, conflicts, polymorphism, alt. splicing)
- Annotation (manual)
- sequence analysis tools
- literature
- external databases (MIM, MGD....)
- Brain storming

Once in Swiss-Prot, the entry is removed from TrEMBL, but remains in EMBL (archive)

CDS: proposed and submitted at EMBL by authors or by genome projects (experimentally proved or derived from gene prediction programs). TrEMBL does not translate DNA sequences, nor use gene prediction programs: only takes CDS given in the EMBL entry.

Archives and databases

- DNA sequence archives
  - EMBL/GenBank/DDBJ is an archive
    - All submitted data goes into the archive
    - Submitters are responsible for the submitted sequences and the accompanying annotation
    - Nobody else can change them (including the curators at EMBL/GenBank/DDBJ)
- Amino acid sequence databases
  - Swiss-Prot is NOT an archive
    - Swiss-Prot chooses what goes into the database and where to place it
    - Swiss-Prot updates annotations and sequences when necessary

From EMBL to TrEMBL

- When submitting a nucleotide sequence, the submitter can specify a CoDing Sequence (CDS).
- This CDS is translated by EMBL/GenBank/DDBJ.
- The corresponding protein sequence is extracted automatically and a TrEMBL entry is created. It contains:
  - the data originally found in the EMBL entry (authors, taxonomy, proposed gene/protein name…)
  - automatic annotation (i.e.domain detection, family attribution…)
  - several additional crosslinks

TrEMBL - the automatically annotated section of UniProtKB

- Rule-based automatic annotation (based on InterPro matches, and machine learning algorithms)
- EVIDENCE TAGS are added to any part of a TrEMBL entry, indicating the source (e.g. import from other dbs, automatic annotations, manual fixes); available as part of the XML format.
- Swiss-Prot evidence tags probably available early 2007

From EMBL to TrEMBL

EBI Dbfetch

```
ID   AY374507   standard; mRNA; VRT; 643 BP.
XX
AC   AY374507;
XX
SV   AY374507.1
XX
DT   23-DEC-2003 (Rel. 78, Created)
DT   10-DEC-2004 (Rel. 82, Last updated, Version 4)
XX
DE   Tetraodon nigroviridis erythropoietin (EPO) mRNA, complete cds.
XX
KW   .
XX
OS   Tetraodon nigroviridis
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Actinopterygii; Neopterygii; Teleostei; Euteleostei; Neoteleostei;
OC   Acanthomorpha; Acanthopterygii; Percomorpha; Tetraodontiformes;
OC   Tetradontoidea; Tetraodontidae; Tetraodon.
XX
RN   [1]
RP   1-643
RX   DOI; 10.1038/nature03025
RX   PUBMED; 15496914.
RA   Jaillon O., Aury J.-M., Brunet F., Petit J.-L., Stange-Thomann N.,
RA   Mauceli E., Bouneau L., Fischer C., Ozouf-Costaz C., Bernot A., Nicaud S.,
RA   Jaffe D., Fisher S., Lutfalla G., Dossat C., Segurens B., Dasilva C.,
RA   Salanoubat M., Levy M., Boudet N., Castellano S., Anthouard V., Jubin C.,
RA   Castelli V., Katinka M., Vacherie B., Biimont C., Skalli Z., Cattolico L.,
RA   Poulain J., Berardinis Vd., Cruaud C., Duprat S., Brottier P.,
RA   Coutanceau J.-P., Gouzy J., Parra G., Lardier G., Chapple C.,
RA   McKernan K.J., McEwan P., Bosak S., Kellis M., Volff J.-N., Guigs R.,
RA   Zody M.C., Mesirov J., Lindblad-Toh K., Birren B., Nusbaum C., Kahn D.,
RA   Robinson-Rechavi M., Laudet V., Schachter V., Quitier F., Saurin W.,
RA   Scarpelli C., Wincker P., Lander E.S., Weissenbach J., Crollius H.R.;
RT   "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the
RT   early vertebrate proto-karyotype";
RL   Nature 431(7011):946-957(2004).
XX
RN   [2]
RP   1-643
RA   Lutfalla G.;
RT   ;
RL   Submitted (26-AUG-2003) to the EMBL/GenBank/DDBJ databases.
RL   UMR5124, CNRS, 1919 route de Mende, Montpellier Cedex 5 34293, France
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..643
FT                   /db_xref="taxon:99883"
FT                   /mol_type="mRNA"
FT                   /organism="Tetraodon nigroviridis"
FT   gene            1..643
FT                   /note="TnEPO"
FT                   /gene="EPO"
FT   CDS             38..625
FT                   /codon_start=1
FT                   /db_xref="GOA:Q6UAM1"
FT                   /db_xref="InterPro:IPR001323"
FT                   /db_xref="InterPro:IPR003013"
FT                   /db_xref="UniProt/TrEMBL:Q6UAM1"
FT                   /note="class I helical cytokine"
FT                   /gene="EPO"
FT                   /product="erythropoietin"
FT                   /protein_id="AAR25698.1"
FT                   /translation="MGLLMCVCVVVFVSGMTGLLAFLLIVLEWTRPSLPSPLRPICDLR
FT                   VLNHFIKEAQDAEEAAMKTCREGCTLSESVVVPQTTVDFDVWEKKNASAKAEEVQSGLWL
FT                   LQQAFNFLRTSVTNAALHSHIDNAVRNLLSVNAVLRSLNIQEFTPQANGAEIEGTWRAS
FT                   SAAELLQVYVNFLRGKARLLLLDAQACQQDVS"
XX
SQ   Sequence 643 BP; 150 A; 177 C; 179 G; 137 T; 0 other;
     ataaagtaaa tcaaagagat gtagtaagag agatgtaatg ggattgttaa tgtgtgtttg        60
     tgttgttgtg tttgtttctg ggatgacagg attgcttgcc ttcctgttga tcgtgttgga       120
     gtggacccgg ccaagcctgc catcgccect gagacccatc tgtgacctga gagtcctcaa       180
     ccatttcatc aaggaggccc aagacgccga agccgccatg aagacctgta gagagggatg       240
     caccctgtca gagtctgtgg tcgttcccca gacaacagtg gattttgacg tctgggagaa       300
     gaaaaacgca tcagcgaagg ccgaggaagt gcagtctggc ttgtggctcc tacagcaggc       360
     cttcaacttc ctacggactt ctgtcaccaa cgcagccctg cacagccaca tagacaacgc       420
     cgtccgaaac ctgctcagcg tcaacgctgt gctgcgcagc ctcaacatcc aggaatttac       480
     cccacaagcc aacgggggcg agatagaggg aacgtggcgg gcgtcctcgg cagcggagct       540
     gctccaggtc tacgtcaact tcctgcgagg caaagcgcgc ctcctcctcc tggacgcgca       600
     ggcctgtcag caagatgtca gctgatgccg gctgacattt cca                         643
//
```

CDS

UniProtKB/TrEMBL entry Q6UAM1

Printer-friendly view | Request update | Quick BlastP search

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

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

Entry information	
Entry name	**Q6UAM1_TETNG**
Primary accession number	**Q6UAM1**
Secondary accession numbers	None
Entered in TrEMBL in	Release 27, July 2004
Sequence was last modified in	Release 27, July 2004
Annotations were last modified in	Release 27, July 2004
Name and origin of the protein	
Protein name	**Erythropoietin**
Synonyms	None
Gene name	**Name: EPO**
From	Tetraodon nigroviridis (Green puffer) [TaxID: 99883]
Taxonomy	Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Actinopterygii, Neopterygii, Teleostei, Euteleostei, Neoteleostei, Acanthomorpha, Acanthopterygii, Percomorpha, Tetraodontiformes, Tetradontoidea, Tetraodontidae, Tetraodon.
References	
[1] NUCLEOTIDE SEQUENCE. DOI=10.1038/nature03025;PubMed=15496914 [NCBI, ExPASy, EBI, Israel, Japan] Jaillon O., Aury J.-M., Brunet F., Petit J.-L., Stange-Thomann N., Mauceli E., Bouneau L., Fischer C., Ozouf-Costaz C., Bernot A., Nicaud S., Jaffe D., Fisher S., Lutfalla G., Dossat C., Segurens B., Dasilva C., Salanoubat M., Levy M., [RSS], Crollius H.R.; "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype."; Nature 431:946-957(2004).	
Comments	
None	
Cross-references	
EMBL	AY374507; AAR25698.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] UniGene Cluster [SIB internal]
Ensembl	AY374507, Tetraodon nigroviridis. [Contig view]
GO	GO:0005576; Cellular component: extracellular region *(inferred from electronic annotation)*. GO:0005128; Molecular function: erythropoietin receptor binding *(inferred from electronic annotation)*. GO:0005179; Molecular function: hormone activity *(inferred from electronic annotation)*. QuickGo view.
InterPro	IPR001323; EPO_TPO. IPR003013; Erythroptn. Graphical view of domain structure.
Pfam	PF00758; EPO_TPO; 1. Pfam graphical view of domain structure.
PIRSF	PIRSF001951; EPO; 1.
PRINTS	PR00272; ERYTHROPTN.
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
HOVERGEN	[Family / Alignment / Tree]
ProtoMap	Q6UAM1.
PRESAGE	Q6UAM1.
ModBase	Q6UAM1.
SMR	Q6UAM1; DA9ECE95B30EB1D9.
SWISS-2DPAGE	Get region on 2D PAGE.
UniRef	View cluster of proteins with at least 50% / 90% identity.
Keywords	
None	
Features	
None	
Sequence information	

Length: **195 AA** | Molecular weight: **21487 Da** | CRC64: **DA9ECE95B30EB1D9** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MGLLMCVCVV VFVSGMTGLL AFLLIVLEWT RPSLPSPLRP ICDLRVLNHF IKEAQDAEAA

        70         80         90        100        110        120
MKTCREGCTL SESVVVPQTT VDFDVWEKKN ASAKAEEVQS GLWLLQQAFN FLRTSVTNAA

       130        140        150        160        170        180
LHSHIDNAVR NLLSVNAVLR SLNIQEFTPQ ANGAEIEGTW RASSAAELLQ VYVNFLRGKA

       190
RLLLLDAQAC QQDVS
```

Q6UAM1 in FASTA format

! TrEMBL neither translates DNA sequences, nor does it use gene prediction programs: **only takes the existing CDS proposed by the submitting authors in the EMBL/Genbank/DDBJ entry**

In particular, the proposed CDS and derived protein sequences can be experimentally proven or derived from gene prediction programs
(this is not obvious from the TrEMBL entry)

TrEMBL does not validate any sequences

From TrEMBL to Swiss-Prot

UniProtKB/TrEMBL entry Q6UAM1

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

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

Entry information	
Entry name	Q6UAM1_TETNG
Primary accession number	Q6UAM1
Secondary accession numbers	None
Entered in TrEMBL in	Release 27, July 2004
Sequence was last modified in	Release 27, July 2004
Annotations were last modified in	Release 27, July 2004
Name and origin of the protein	
Protein name	**Erythropoietin**
Synonyms	None
Gene name	**Name: EPO**
From	Tetraodon nigroviridis (Green puffer) [TaxID: 99883]
Taxonomy	Eukaryota, Metazoa, Chordata, Craniata, Vertebrata, Euteleostomi, Actinopterygii, Neopterygii, Teleostei, Tetraodontiformes, Tetraodontoidea, Tetraodontidae, Tetraodon
References	
[1] NUCLEOTIDE SEQUENCE. DOI=10.1038/nature03025; PubMed=15496914 [NCBI, EuPASy, EBI, Israel, Japan] Jaillon O., Aury J.-M., Brunet F., Petit J.-L., Stange-Thomann N., Mauceli E., Bouneau L., Fischer C., Ozouf-Costaz C., Bernot A., Nicaud S., Jaffe D., Fisher S., Lutfalla G., Dossat C., Segurens B., Dasilva C., Salanoubat M., Levy M., Boudet N., Castellano S., Anthouard V., Jubin C., Castelli V., Katinka M., Vacherie B., Biemont C., Skalli Z., Cattolico L., Poulain J., De Berardinis V., Cruaud C., Duprat S., Brottier P., Coutanceau J.-P., Gouzy J., Parra G., Lardier G., Chapple C., McKernan K.J., McEwan P., Bosak S., Kellis M., Volff J.-N., Guigo R., Zody M.C., Mesirov J., Lindblad-Toh K., Birren B., Nusbaum C., Kahn D., Robinson-Rechavi M., Laudet V., Schachter V., Quetier F., Saurin W., Scarpelli C., Wincker P., Lander E.S., Weissenbach J., Roest Crollius H.R.; "Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype."; Nature 431:946-957(2004)	
Comments	
None	
Cross-references	
EMBL	AY374507; AAR25698.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] UniGene Cluster [SIB internal]
Ensembl	AY374507; Tetraodon nigroviridis. [Contig view]
GO	GO:0005576; Cellular component: extracellular region *(inferred from electronic annotation)*. GO:0005128; Molecular function: erythropoietin receptor binding *(inferred from electronic annotation)*. GO:0005179; Molecular function: hormone activity *(inferred from electronic annotation)*. QuickGo view
InterPro	IPR001323; EPO_TPO. IPR003013; Erythroptn. Graphical view of domain structure.
Pfam	PF00758; EPO_TPO; 1. Pfam graphical view of domain structure.
PIRSF	PIRSF001951; EPO; 1.
PRINTS	PR00272; ERYTHROPTN.
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
HOVERGEN	[Family / Alignment / Tree]
ProtoMap	Q6UAM1.
PRESAGE	Q6UAM1.
ModBase	Q6UAM1.
SMR	Q6UAM1; DA9ECE95B30EB1D9.
SWISS-2DPAGE	Get region on 2D PAGE.
UniRef	View cluster of proteins with at least 50% / 90% identity.
Keywords	
None	
Features	
None	
Sequence information	
Length: **195 AA**	Molecular weight: **21487 Da** CRC64: **DA9ECE95B30EB1D9** [This is a checksum on the sequence]

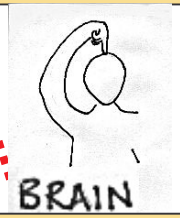
```
        10         20         30         40         50         60
MGLLBCVCVV VFVSGNTGLE AFLLIVLEHT RPSLPSPLEP ICDLRVLSRF IKEAQDAEAA

        70         80         90        100        110        120
EKTCRRGCTL SEKVVVVPQT T VDPDVMEKDN ASAKAEEVQE GLMLLQQATH FLKTDVTHAA

       130        140        150        160        170        180
LBSNISVAVK NLLSVNAVLE SLNIQEFTPQ ANGAEIKCTR RADSAAELLQ YTVNFLRGKA

       190
RLLLLPAQAC QQDVS
```

Q6UAM1 in FASTA format

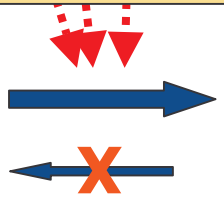
Literature information
(more than 1500 journals cited)

Sequence analysis
High performance bioinformatics tools

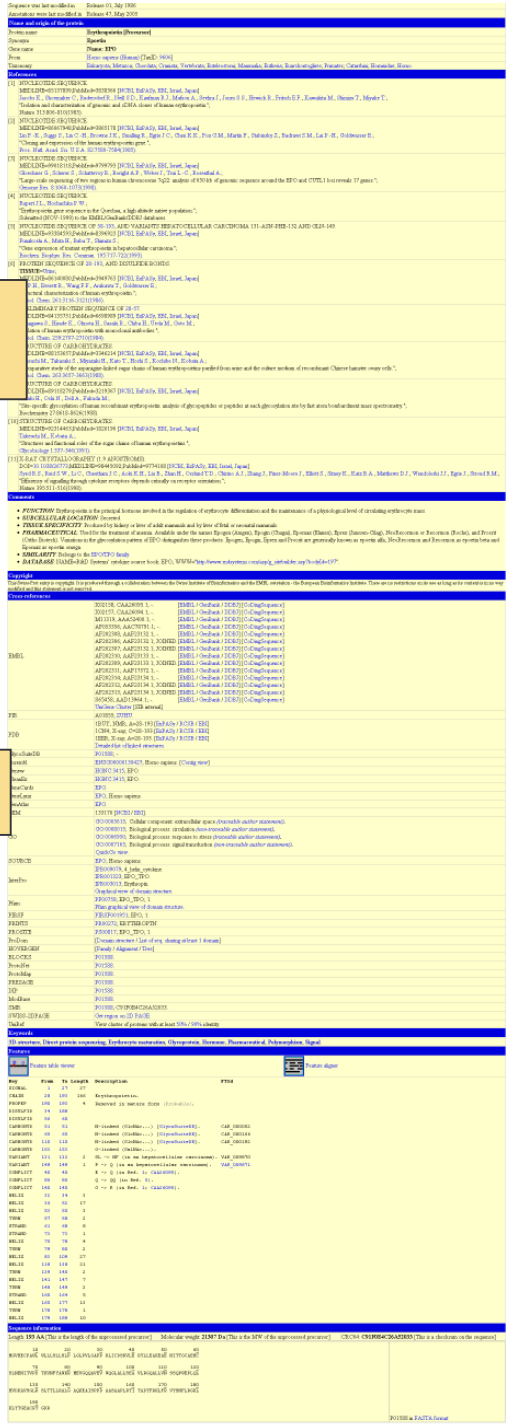
Databases and
external scientific expertise

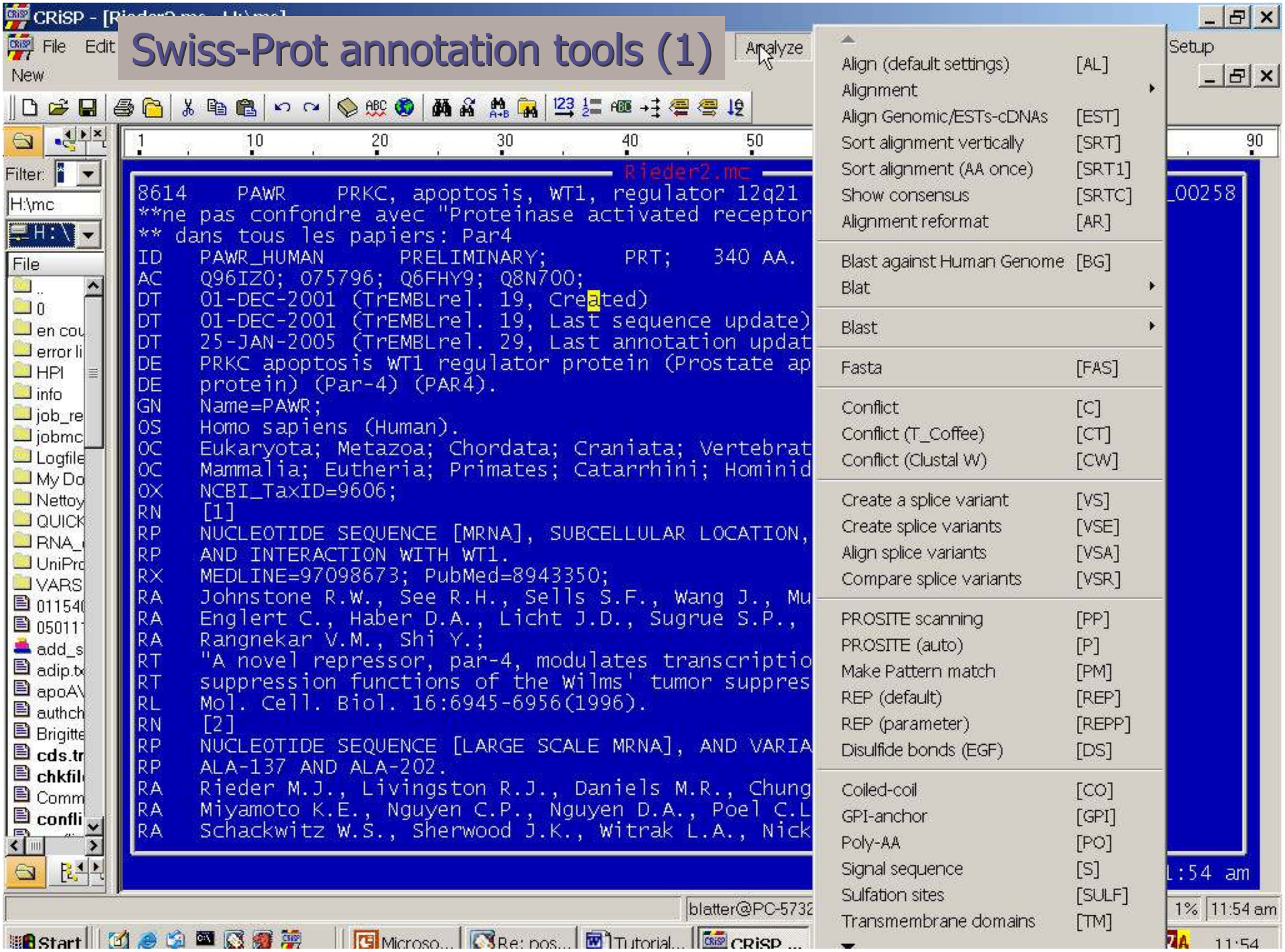


Annotation and sequence check

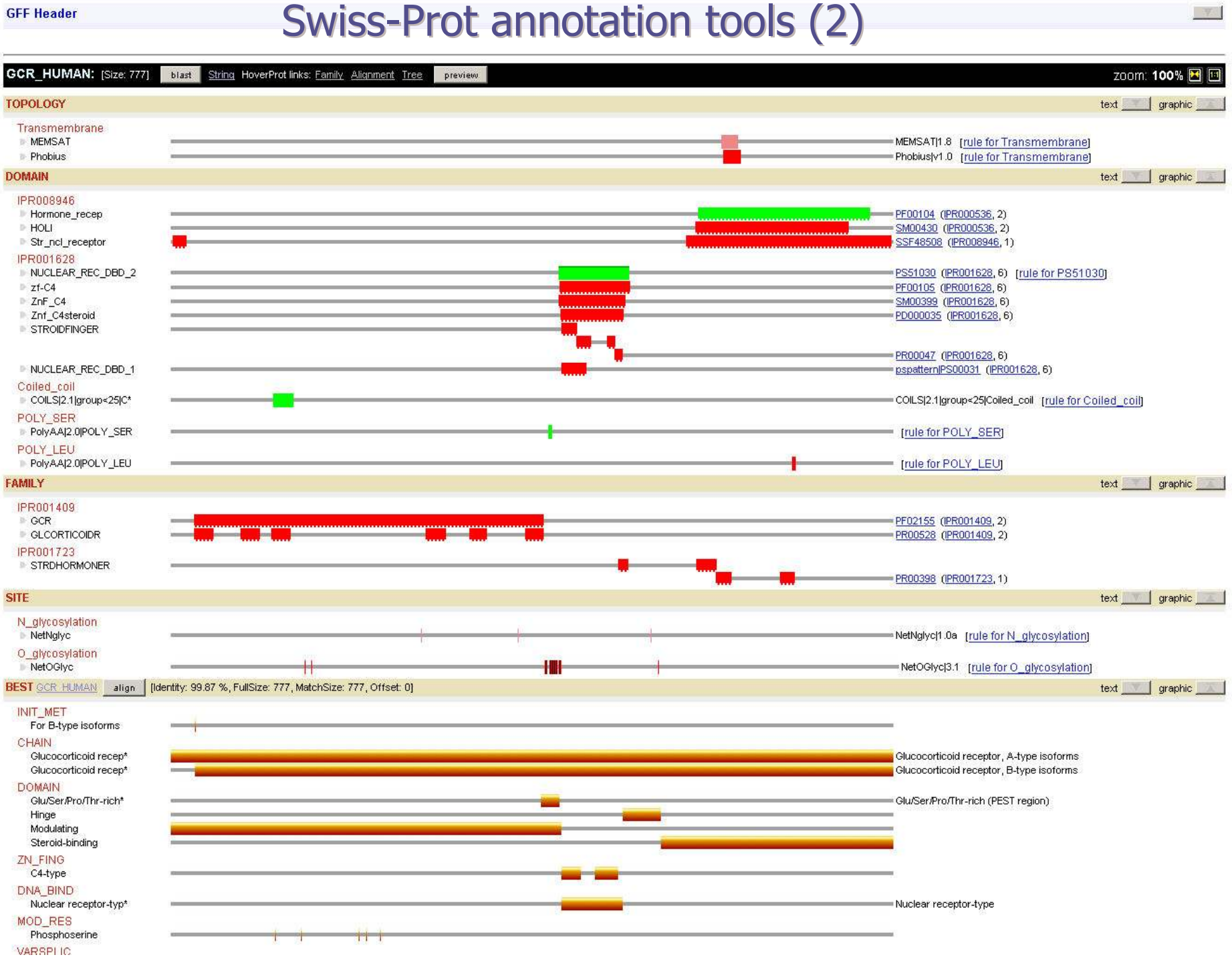


Once manually annotated and integrated into Swiss-Prot, the entry is deleted from TrEMBL

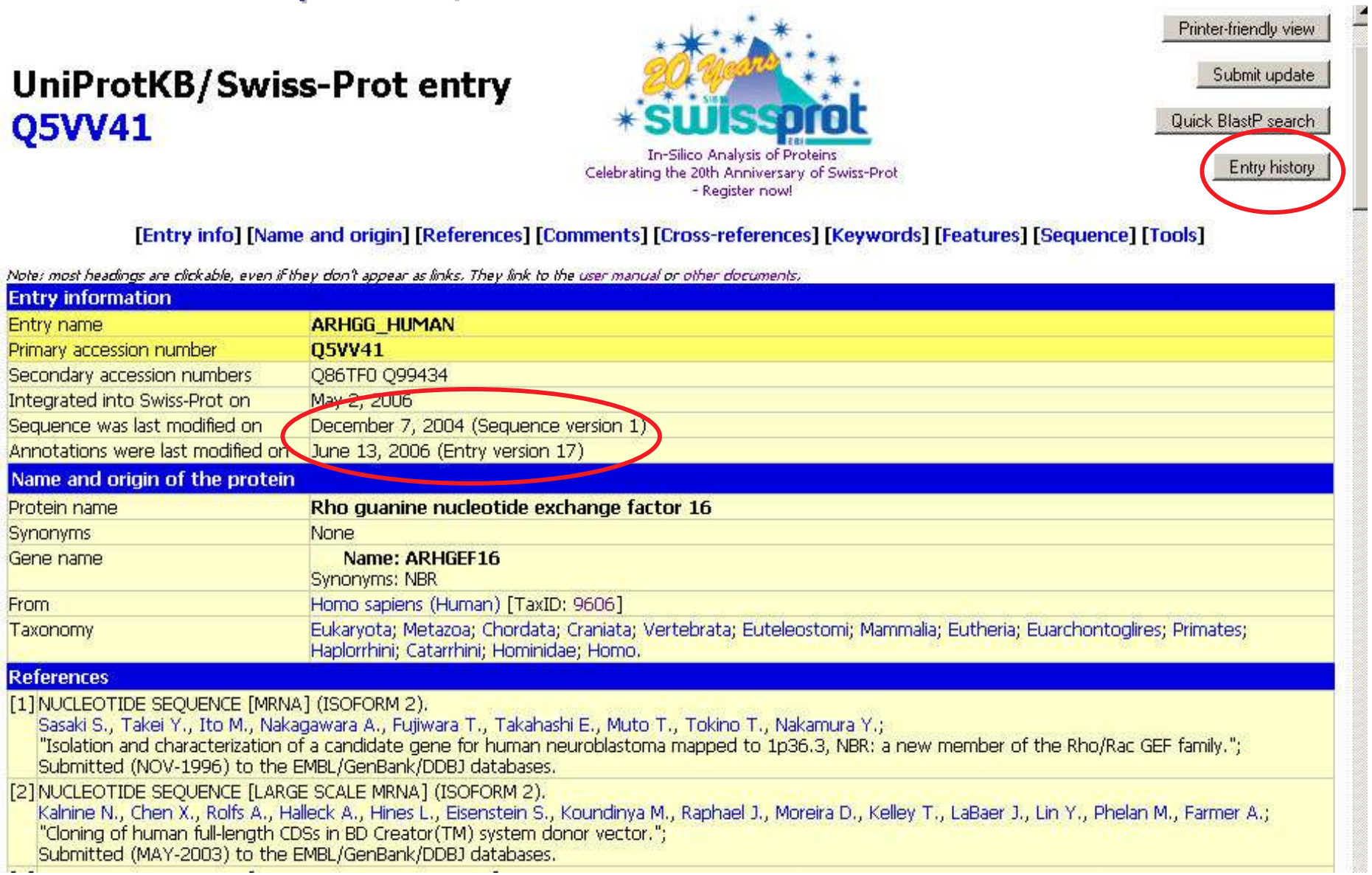



Swiss-Prot annotation tools (1)
CRiSP - [Rieder2.mc - H:\mc]
File Edit
New
Analyze
Setup
Filter:
H:\mc
H:\
File
Rieder2.mc
8614 PAWR PRKC, apoptosis, WT1, regulator 12q21
**ne pas confondre avec "Proteinase activated receptor
** dans tous les papiers: Par4
ID PAWR_HUMAN PRELIMINARY; PRT; 340 AA.
AC Q96IZ0; O75796; Q6FHY9; Q8N700;
DT 01-DEC-2001 (TrEMBLrel. 19, Created)
DT 01-DEC-2001 (TrEMBLrel. 19, Last sequence update)
DT 25-JAN-2005 (TrEMBLrel. 29, Last annotation updat
DE PRKC apoptosis WT1 regulator protein (Prostate ap
DE protein) (Par-4) (PAR4).
GN Name=PAWR;
OS Homo sapiens (Human).
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrat
OC Mammalia; Eutheria; Primates; Catarrhini; Hominid
OX NCBI_TaxID=9606;
RN [1]
RP NUCLEOTIDE SEQUENCE [MRNA], SUBCELLULAR LOCATION,
RP AND INTERACTION WITH WT1.
RX MEDLINE=97098673; PubMed=8943350;
RA Johnstone R.W., See R.H., Sells S.F., Wang J., Mu
RA Englert C., Haber D.A., Licht J.D., Sugrue S.P.,
RA Rangnekar V.M., Shi Y.;
RT "A novel repressor, par-4, modulates transcriptio
RT suppression functions of the Wilms' tumor suppres
RL Mol. Cell. Biol. 16:6945-6956(1996).
RN [2]
RP NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA], AND VARIA
RP ALA-137 AND ALA-202.
RA Rieder M.J., Livingston R.J., Daniels M.R., Chung
RA Miyamoto K.E., Nguyen C.P., Nguyen D.A., Poel C.L
RA Schackwitz W.S., Sherwood J.K., Witrak L.A., Nick
Align (default settings) [AL]
Alignment
Align Genomic/ESTs-cDNAs [EST]
Sort alignment vertically [SRT]
Sort alignment (AA once) [SRT1]
Show consensus [SRTC]
Alignment reformat [AR]
Blast against Human Genome [BG]
Blat
Blast
Fasta [FAS]
Conflict [C]
Conflict (T_Coffee) [CT]
Conflict (Clustal W) [CW]
Create a splice variant [VS]
Create splice variants [VSE]
Align splice variants [VSA]
Compare splice variants [VSR]
PROSITE scanning [PP]
PROSITE (auto) [P]
Make Pattern match [PM]
REP (default) [REP]
REP (parameter) [REPP]
Disulfide bonds (EGF) [DS]
Coiled-coil [CO]
GPI-anchor [GPI]
Poly-AA [PO]
Signal sequence [S]
Sulfation sites [SULF]
Transmembrane domains [TM]
blatter@PC-5732
Start
11:54 am

Swiss-Prot annotation tools (2)



Monitoring entry history: The UniProtKB Sequence/Annotation Version archive



UniProtKB/Swiss-Prot entry
Q5VV41

In-Silico Analysis of Proteins
Celebrating the 20th Anniversary of Swiss-Prot
- Register now!

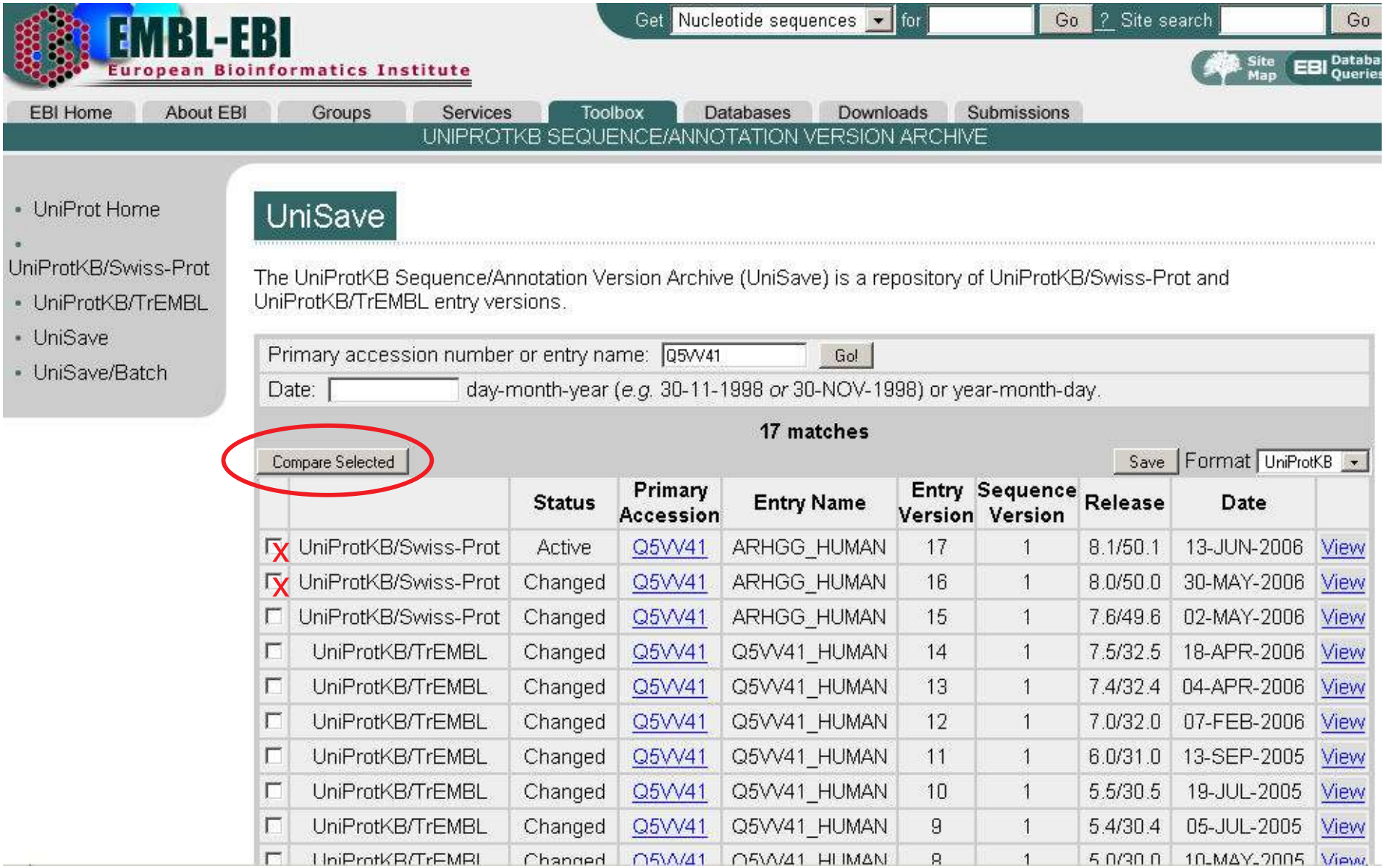
Printer-friendly view
Submit update
Quick BlastP search
Entry history

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

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

Entry information	
Entry name	**ARHGG_HUMAN**
Primary accession number	**Q5VV41**
Secondary accession numbers	Q86TF0 Q99434
Integrated into Swiss-Prot on	May 2, 2006
Sequence was last modified on	December 7, 2004 (Sequence version 1)
Annotations were last modified on	June 13, 2006 (Entry version 17)
Name and origin of the protein	
Protein name	**Rho guanine nucleotide exchange factor 16**
Synonyms	None
Gene name	**Name: ARHGEF16** Synonyms: NBR
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
References	
[1]	NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 2). Sasaki S., Takei Y., Ito M., Nakagawara A., Fujiwara T., Takahashi E., Muto T., Tokino T., Nakamura Y.; "Isolation and characterization of a candidate gene for human neuroblastoma mapped to 1p36.3, NBR: a new member of the Rho/Rac GEF family."; Submitted (NOV-1996) to the EMBL/GenBank/DDBJ databases.
[2]	NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORM 2). Kalnine N., Chen X., Rolfs A., Halleck A., Hines L., Eisenstein S., Koundinya M., Raphael J., Moreira D., Kelley T., LaBaer J., Lin Y., Phelan M., Farmer A.; "Cloning of human full-length CDSs in BD Creator(TM) system donor vector."; Submitted (MAY-2003) to the EMBL/GenBank/DDBJ databases.

Monitoring entry history: The UniProtKB Sequence/Annotation Version archive



	Status	Primary Accession	Entry Name	Entry Version	Sequence Version	Release	Date	
X	UniProtKB/Swiss-Prot	Active	Q5VV41	ARHGG_HUMAN	17	1	8.1/50.1	13-JUN-2006	View
X	UniProtKB/Swiss-Prot	Changed	Q5VV41	ARHGG_HUMAN	16	1	8.0/50.0	30-MAY-2006	View
	UniProtKB/Swiss-Prot	Changed	Q5VV41	ARHGG_HUMAN	15	1	7.6/49.6	02-MAY-2006	View
	UniProtKB/TrEMBL	Changed	Q5VV41	Q5VV41_HUMAN	14	1	7.5/32.5	18-APR-2006	View
	UniProtKB/TrEMBL	Changed	Q5VV41	Q5VV41_HUMAN	13	1	7.4/32.4	04-APR-2006	View
	UniProtKB/TrEMBL	Changed	Q5VV41	Q5VV41_HUMAN	12	1	7.0/32.0	07-FEB-2006	View
	UniProtKB/TrEMBL	Changed	Q5VV41	Q5VV41_HUMAN	11	1	6.0/31.0	13-SEP-2005	View
	UniProtKB/TrEMBL	Changed	Q5VV41	Q5VV41_HUMAN	10	1	5.5/30.5	19-JUL-2005	View
	UniProtKB/TrEMBL	Changed	Q5VV41	Q5VV41_HUMAN	9	1	5.4/30.4	05-JUL-2005	View
	UniProtKB/TrEMBL	Changed	Q5VV41	Q5VV41_HUMAN	8	1	5.0/30.0	10-MAY-2005	View

UniSave

The UniProtKB Sequence/Annotation Version Archive (UniSave) is a repository of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions.

Primary accession number or entry name: Q5VV41 Go!

Date: day-month-year (*e.g.* 30-11-1998 *or* 30-NOV-1998) or year-month-day.

<< Earlier Later >> Q5VV41 30-MAY-2006

<< Earlier Q5VV41 13-JUN-2006

Back to List

```
ID   ARHGG_HUMAN     STANDARD;       PRT;    709 AA.
AC   Q5VV41; Q86TF0; Q99434;
DT   02-MAY-2006, integrated into UniProtKB/Swiss-Prot.
DT   07-DEC-2004, sequence version 1.
DT   13-JUN-2006, entry version 17.
DT   30-MAY-2006, entry version 16.
DE   Rho guanine nucleotide exchange factor 16.
GN   Name=ARHGEF16; Synonyms=NBR;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Verte
```

Human chromosome 1 reference published

New database cross-reference added

```
DR   Ensembl; ENSG00000130762; Homo sapiens.
DR   HGNC; HGNC:15515; ARHGEF16.
DR   RZPD-ProtExp; IOH27071; -.
DR   RZPD-ProtExp; IOH5304; -.
DR   RZPD-ProtExp; T0410; -.
DR   InterPro; IPR001849; PH.
DR   InterPro; IPR000219; RhoGEF.
DR   InterPro; IPR001452; SH3.
DR   Pfam; PF00169; PH; 1.
```

```
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
RX   PubMed=16710414; DOI=10.1038/nature04727;
RA   Gregory S.G., Barlow K.F., McLay K.E., Kaul R., Swarbreck D.,
RA   Dunham A., Scott C.E., Howe K.L., Woodfine K., Spencer C.C.A.,
RA   Jones M.C., Gillson C., Searle S., Zhou Y., Kokocinski F.,
RA   McDonald L., Evans R., Phillips K., Atkinson A., Cooper R., Jones C.,
RA   Hall R.E., Andrews T.D., Lloyd C., Ainscough R., Almeida J.P.,
RA   Ambrose K.D., Anderson F., Andrew R.W., Ashwell R.I.S., Aubin K.,
RA   Babbage A.K., Bagguley C.L., Bailey J., Beasley H., Bethel G.,
RA   Bird C.P., Bray-Allen S., Brown J.Y., Brown A.J., Buckley D.,
RA   Burton J., Bye J., Carder C., Chapman J.C., Clark S.Y., Clarke G.,
RA   Clee C., Cobley V., Collier R.E., Corby N., Coville G.J., Davies J.,
RA   Deadman R., Dunn M., Earthrowl M., Ellington A.G., Errington H.,
RA   Frankish A., Frankland J., French L., Garner P., Garnett J., Gay L.,
RA   Ghori M.R.J., Gibson R., Gilby L.M., Gillett W., Glithero R.J.,
RA   Grafham D.V., Griffiths C., Griffiths-Jones S., Grocock R.,
RA   Hammond S., Harrison E.S.I., Hart E., Haugen E., Heath P.D.,
RA   Holmes S., Holt K., Howden P.J., Hunt A.R., Hunt S.E., Hunter G.,
RA   Isherwood J., James R., Johnson C., Johnson D., Joy A., Kay M.,
RA   Kershaw J.K., Kibukawa M., Kimberley A.M., King A., Knights A.J.,
RA   Lad H., Laird G., Lawlor S., Leongamornlert D.A., Lloyd D.M.,
RA   Loveland J., Lovell J., Lush M.J., Lyne R., Martin S.,
RA   Mashreghi-Mohammadi M., Matthews L., Matthews N.S.W., McLaren S.,
RA   Milne S., Mistry S., Moore M.J.F., Nickerson T., O'Dell C.N.,
RA   Oliver K., Palmeiri A., Palmer S.A., Parker A., Patel D., Pearce A.V.,
RA   Peck A.I., Pelan S., Phelps K., Phillimore B.J., Plumb R., Rajan J.,
RA   Raymond C., Rouse G., Saenphimmachak C., Sehra H.K., Sheridan E.,
RA   Shownkeen R., Sims S., Skuce C.D., Smith M., Steward C.,
RA   Subramanian S., Sycamore N., Tracey A., Tromans A., Van Helmond Z.,
RA   Wall M., Wallis J.M., White S., Whitehead S.L., Wilkinson J.E.,
RA   Willey D.L., Williams H., Wilming L., Wray P.W., Wu Z., Coulson A.,
RA   Vaudin M., Sulston J.E., Durbin R., Hubbard T., Wooster R., Dunham I.,
RA   Carter N.P., McVean G., Ross M.T., Harrow J., Olson M.V., Beck S.,
RA   Rogers J., Bentley D.R.;
RT   "The DNA sequence and biological annotation of human chromosome 1.";
RL   Nature 441:315-321(2006).
RG   Human chromosome 1 international sequencing consortium;
RL   Submitted (MAY-2005) to the EMBL/GenBank/DDBJ databases.
RN   [4]
```

Direct access to the UniProtKB Sequence/Annotation Version archive (UniSave):

http://www.ebi.ac.uk/uniprot/unisave/

Access to the UniProt Knowledgebase

Direct access (keyword search)

- Sequence Retrieval System (SRS, Europe)
- Entrez (NCBI, USA) – Swiss-Prot (not TrEMBL) is integrated in GenPept, but with a changed format, and with some information (e.g. cross-references) removed
- Query tools on ExPASy & UniProt (http://www.expasy.org/sprot/, http://www.uniprot.org)

Indirect access (sequence search)

- Bioinformatics & sequence analysis tools (Blast, Fasta, GCG, Emboss, MS Identification tools…)

Downloading the UniProt Knowledgebase

http://www.expasy.org/sprot/download.html

- Swiss-Prot and TrEMBL form a complete, non-redundant database, the UniProt Knowledgebase
- Can be downloaded from ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase
- In « Swiss-Prot » format, fasta or xml format
- Complemented by sequences of alternative splice isoforms
- « all » about « all » proteins! (at least all CDS submitted to the public nucleotide sequence databases)

FTP access:
ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/complete/

docs		13.06.2006	14:00:00
keydef.xml.gz	94 KB	21.03.2006	15:00:00
reldate.txt	1 KB	13.06.2006	14:00:00
uniprot.dtd.gz	4 KB	13.06.2006	14:00:00
uniprot.xsd	41 KB	13.06.2006	14:00:00
uniprot.xsd.gz	6 KB	13.06.2006	14:00:00
uniprot_sprot.dat.gz	180957 KB	13.06.2006	14:00:00
uniprot_sprot.fasta.gz	35339 KB	13.06.2006	14:00:00
uniprot_sprot.xml.gz	226450 KB	13.06.2006	14:00:00
uniprot_sprot_varsplic.fasta.gz	3119 KB	13.06.2006	14:00:00
uniprot_trembl.dat.gz	1126865 KB	13.06.2006	14:00:00
uniprot_trembl.fasta.gz	555389 KB	13.06.2006	14:00:00
uniprot_trembl.xml.gz	1331922 KB	13.06.2006	14:00:00

Swiss-Prot and TrEMBL are also available on CD-ROM: datalib@ebi.ac.uk

TrEMBL XML includes evidence tags

If you want to develop tools to work with your local copy of Swiss-Prot/TrEMBL:

Swissknife – a PERL parser for UniProtKB

Constantly updated according to latest format changes

Advantage: you do not need to know how exactly the information is stored in the flat file

- http://swissknife.sourceforge.net/
- ftp://ftp.ebi.ac.uk/pub/software/swissprot/Swissknife/

Swiss-Prot & TrEMBL
introduce a new arithmetical concept!

220’000 + 2’000’000 ≈ 1’500’000

~9500 species ~100’000 species

Redundancy in TrEMBL

&

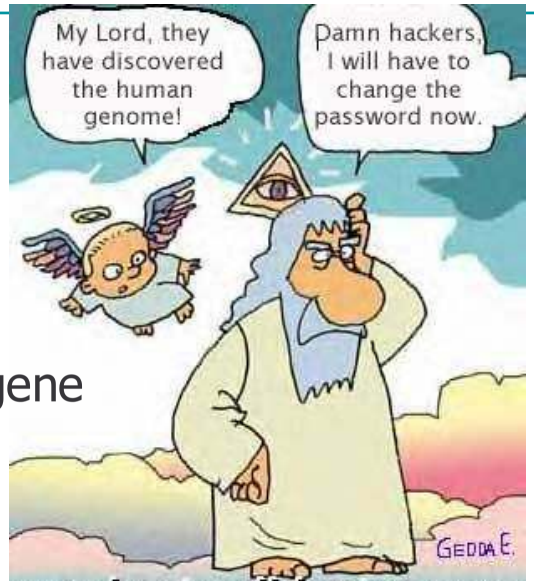
Redundancy between TrEMBL and Swiss-Prot

Only 100% identical sequences (except fragments) from the same organism are merged automatically (in TrEMBL)

All other merging operations are more complex and are performed manually (in Swiss-Prot)

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

14’094 + 55’000 ≈ about 22’000*



* human gene number estimation:
<25’000

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

Missing sequences:

- Sequences not submitted to EMBL/GenBank/DDBJ, except those submitted directly to Swiss-Prot or PIR
- Not yet predicted or known genes ("no CDS provided by the submitters" or no DNA sequence Confidential data (Patent application sequences)
- Immunoglobulins, T-cell receptors (-> UniParc)
- …

Non-Redundant Complete Proteome Sets

- Text search (e.g. SRS): UniProtKB keyword « Complete proteome », combined with an organism name
- Download precomputed sets (bacteria, archaea, some eukaryotes): ftp://ftp.expasy.org/databases/complete_proteomes/entries
- EBI Integr8 http://www.ebi.ac.uk/integr8/

Distinguishing Swiss-Prot and TrEMBL

- Definitions
  - A TrEMBL entry is a computer-annotated record derived from a certain coding sequence (CDS) in the EMBL nucleotide sequence database not in Swiss-Prot, after some redundancy removal and automated annotation.
  - A Swiss-Prot entry is a manually annotated record for a certain protein.

Distinguishing Swiss-Prot and TrEMBL

ID lines

- Data Class

Swiss-Prot ID 1433B_BOVIN STANDARD; PRT; 245 AA.
TrEMBL ID Q28835_RABIT PRELIMINARY; PRT; 90 AA.

- Entry name

Swiss-Prot ID 1433B_BOVIN STANDARD; PRT; 245 AA
TrEMBL ID Q28835_RABIT PRELIMINARY; PRT; 90 AA.

Distinguishing Swiss-Prot and TrEMBL

AC lines

Swiss-Prot AC P29358;

TrEMBL AC Q28835;

(-> no difference! When the TrEMBL entry is annotated and moves to Swiss-Prot, it will keep its accession number)

DT lines

Swiss-Prot

DT 01-JUL-1993, integrated into UniProtKB/Swiss-Prot.

TrEMBL

DT 01-NOV-1996, integrated into UniProtKB/TrEMBL.

TrEMBL <-> Swiss-Prot summary

- Almost all Swiss-Prot entries are derived from TrEMBL entries
- TrEMBL entries are entries waiting to be integrated in Swiss-Prot
- TrEMBL grows faster than Swiss-Prot, so some TrEMBL entries will never get annotated
- TrEMBL contains redundancy
- Swiss-Prot is completely non-redundant
- For human sequences, it takes >4 TrEMBL entries to make one Swiss-Prot entry

Take home message

- Swiss-Prot is the **non redundant**, **manually annotated** and **highly cross-referenced** section of the UniProt Knowledgebase
- Be aware of the differences between UniProtKB/TrEMBL and UniProtKB/Swiss-Prot
  - Computer vs. Human
  - Redundant vs. Non-redundant
- **Always** cite the Accession number, not the ID
  - The AC is stable
  - The ID might change

We need **your** feedback and **your** expertise!

swiss-prot@expasy.org

http://www.expasy.org/sprot/update.html

(and from every Swiss-Prot entry page on ExPASy)

Righting the wrongs

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

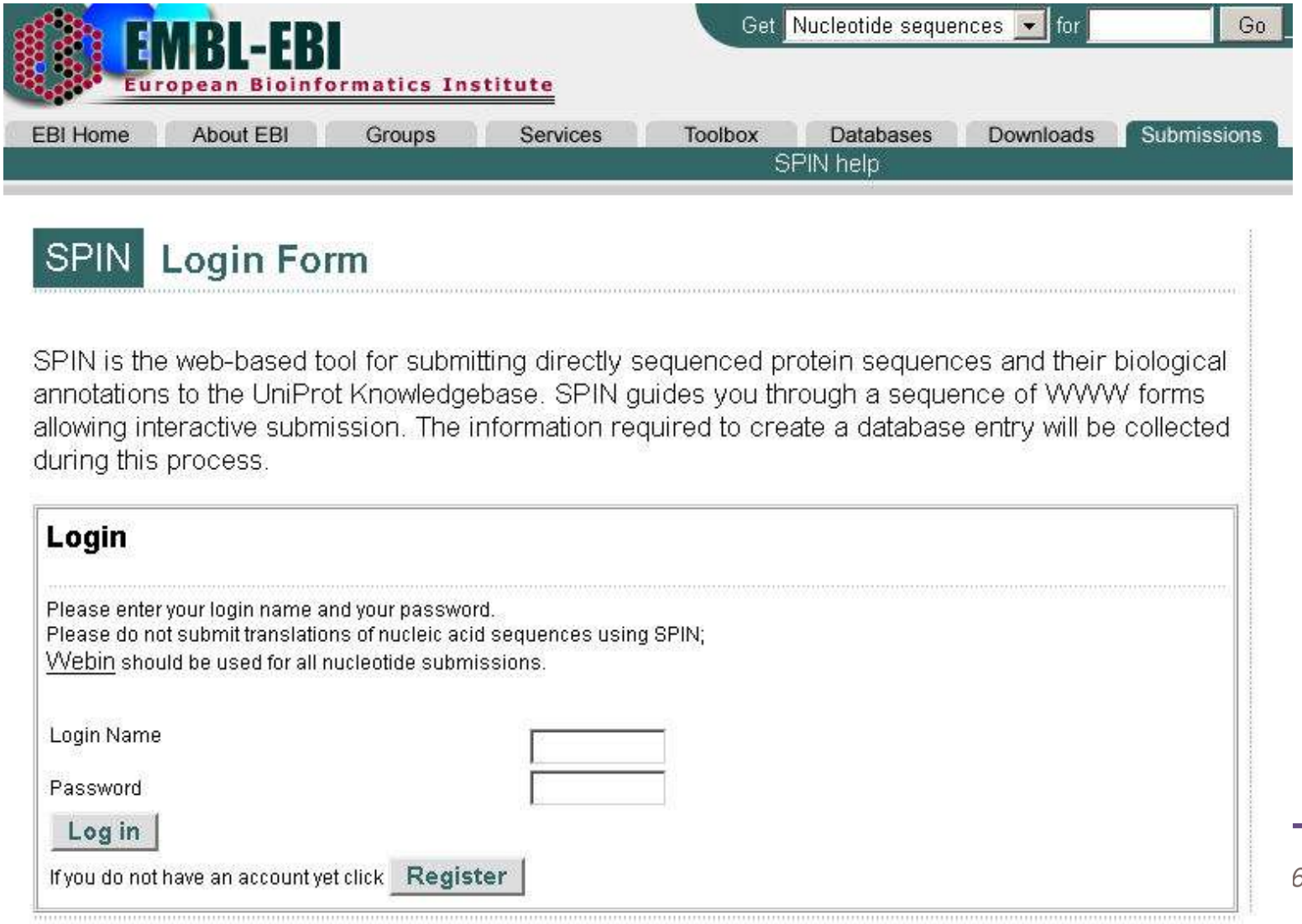
"Sequencing error rates: ~1 base in 10'000"

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

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

Submitting sequences to UniProtKB/Swiss-Prot:

http://www.ebi.ac.uk/swissprot/Submissions/spin/



UniRef100, 90 and 50 clusters

One **UniRef100** entry -> all **identical sequences** from UniProtKB and some sections of UniParc (including fragments, Swiss-Prot splice variants).

One **UniRef90** entry -> sequences that have at least **90% or more identity.**

One **UniRef50** entry -> sequences that are **at least 50% identical**.

UniRef100, 90 and 50 clusters

One cluster can contain sequences of several species, clustering is done independently of the organism

Each cluster has a « representative », « reference » sequence, preferably that of the best-annotated Swiss-Prot entry

UniRef identifiers are of the form UniRef100_P99999, UniRef50_P00414 – not stable, as clusters are recomputed with every biweekly release, and cluster representatives can change!

UniRef is useful for **comprehensive BLAST** sequence searches by providing sets of representative sequences.

BLAST against UniRef www.expasy.org/tools/blast/

*Enter a Swiss-Prot/TrEMBL accession number or a **PROTEIN** sequence in **RAW format**.*

```
TMDKSELVQKAKLAEQAERYDDMAAAMKAVTEQGHELSNEERNLLSVAYKNVVGARRSSW
RVISSIEQKTERNEKKQQMGKEYREKIEAELQDICNDVLQLLDKYLIPNATQPESKVFYL
KMKGDYFRYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFYY
EILNSPEKACSLAKTAFDEAIAELDTLNEESYKDSTLIMQLLRDNLTLWTSENQGDEGDA
GEGEN
```

Output format: HTML

Run BLAST or Reset Form

Choose the appropriate BLAST program and database:

blastp - query against the UniProt knowledgebase (Swiss-Prot + TrEMBL)

Default: UniProt

Taxonomic groups (not available for PDB and translated EST):

select a database subsection: - Complete database -

To restrict the search to a particular taxon, it is much faster to select a database subsection specify your own taxonomic group in the box below. This also gives more accurate statistics

or specify a taxonomic group

Enter a species name, a TaxID or the latin name of a taxonomic group (elements of the OC, particular taxon. You may enter a list separated by semicolons (";"). Example: Fungi; Homo

or select a microbial proteome

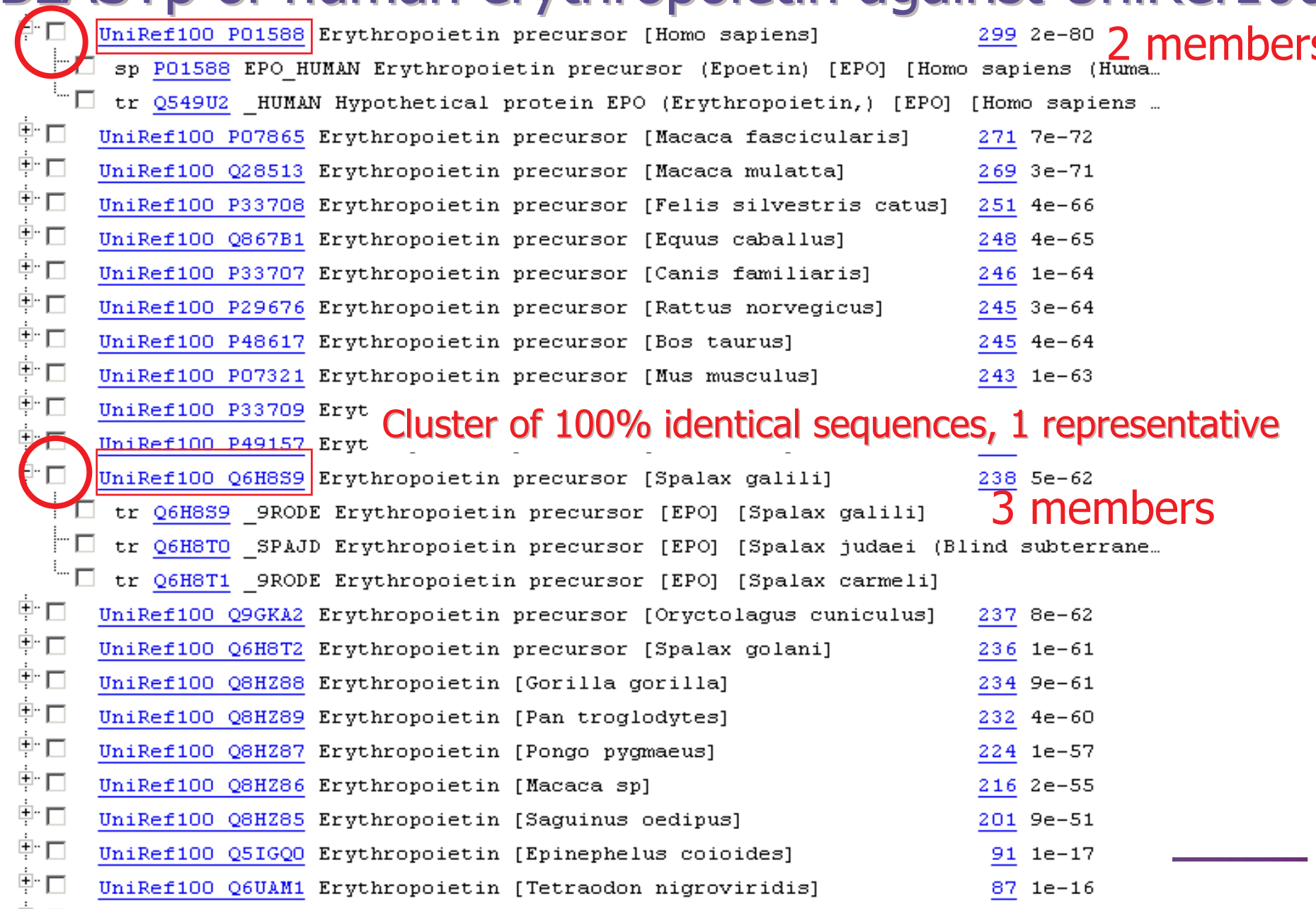
Non-redundant Swis the HAMAP pages.

Search only Swiss-Prot (curated sequences) Exclude fragment sequences

blastp - query against another protein database UniRef50

UniRef100/90/50

BLASTp of human erythropoietin against UniRef100



BLASTp of human erythropoietin against UniRef90

- UniRef90 P01588 Erythropoietin precursor related cluster 299 2e-80
  - sp P01588 EPO_HUMAN Erythropoietin precursor (Epoetin) [EPO] [Homo sapiens (Huma…
  - tr Q549U2 _HUMAN Hypothetical protein EPO (Erythropoietin,) [EPO] [Homo sapiens …
  - sp P07865 EPO_MACFA Erythropoietin precursor [EPO] [Macaca fascicularis (Crab ea…
  - sp Q28513 EPO_MACMU Erythropoietin precursor [EPO] [Macaca mulatta (Rhesus macaq…
  - tr Q8HZ86 _9PRIM Erythropoietin (Fragment) [Macaca sp]
  - tr Q8HZ88 _9PRIM Erythropoietin (Fragment) [Gorilla gorilla (gorilla)]
  - tr Q8HZ89 _PANTR Erythropoietin (Fragment) [Pan troglodytes (Chimpanzee)]
  - tr Q8HZ87 _PONPY Erythropoietin (Fragment) [Pongo pygmaeus (Orangutan)]
- UniRef90 Q867B1 Erythropoietin precursor related cluster 248 2e-65
- UniRef90 P33707 Erythropoietin precursor related cluster 246 9e-65
- UniRef90 P07321 Erythropoietin precursor related cluster 243 8e-64
- UniRef90 P33709 Erythropoietin precursor related cluster 240 7e-63
- UniRef90 P49157 Erythropoietin precursor related cluster 239 2e-62
- UniRef90 Q6H8S9 Erythropoietin precursor related cluster 238 3e-62
- UniRef90 Q9GKA2 Erythropoietin precursor related cluster 237 6e-62
- UniRef90 Q8HZ85 Erythropoietin related cluster 201 6e-51
- UniRef90 Q5IGQ0 Erythropoietin related cluster 91 7e-18
- UniRef90 Q6UAM1 Erythropoietin related cluster 87 1e-16
- UniRef90 Q6JV22 Erythropoietin related cluster 84 8e-16
- UniRef90 Q9QV40 Erythropoietin related cluster 67 2e-10

8 members

BLASTp of human erythropoietin against UniRef50

22 members!

UniRef50_P01588	Erythropoietin precursor related cluster	299	1e-80
UniRef50_Q6UAM1	Erythropoietin related cluster	87	6e-17
UniRef50_Q9QV40	Erythropoietin related cluster	67	1e-10
UniRef50_Q6IYE9	Thrombopoietin related cluster	40	0.008
UniRef50_P37024	ATP-dependent helicase hrpB related cluster	31	5.0
UniRef50_Q8PBS6	Hypothetical protein XCC1043 related cluster	30	8.5
UniRef50_Q4ZT67	Amino acid adenylation related cluster	30	8.5

UniRef combines closely related sequences into a single record to speed sequence searches. Such entries are indicated in the match list below with the '+' sign. Click on it to display the related sequences.

List of potentially matching sequences

Send selected sequences to: Clustal W (multiple alignment) | Submit Query | Select up to...

- Clustal W (multiple alignment)
- T-COFFEE (multiple alignment)
- Reduce redundancy
- PRATT (find conserved patterns)
- Retrieve entries (Swiss-Prot format)
- Retrieve sequences (FASTA format)
- Retrieve matching entries accession codes

Include query sequence

DbAC

Submission of selected sequences to further analysis

UniRef90 P01588 Erythropoietin precursor related cluster 299 2e-80
- sp P01588 EPO_HUMAN Erythropoietin precursor (Epoetin) [EPO] [Homo sapiens (Huma…
- tr Q549U2 _HUMAN Hypothetical protein EPO (Erythropoietin,) [EPO] [Homo sapiens …
- sp P07865 EPO_MACFA Erythropoietin precursor [EPO] [Macaca fascicularis (Crab ea…
- sp Q28513 EPO_MACMU Erythropoietin precursor [EPO] [Macaca mulatta (Rhesus macaq…
- tr Q8HZ86 _9PRIM Erythropoietin (Fragment) [Macaca sp]
- tr Q8HZ88 _9PRIM Erythropoietin (Fragment) [Gorilla gorilla (gorilla)]
- tr Q8HZ89 _PANTR Erythropoietin (Fragment) [Pan troglodytes (Chimpanzee)]
- tr Q8HZ87 _PONPY Erythropoietin (Fragment) [Pongo pygmaeus (Orangutan)]

UniRef90 Q867B1 Erythropoietin precursor related cluster 248 2e-65

UniRef90 P33707 Erythropoietin precursor related cluster 246 9e-65

UniRef90 P07321 Erythropoietin precursor related cluster 243 8e-64

UniRef90 P33709 Erythropoietin precursor related cluster 240 7e-63

UniRef90 P49157 Erythropoietin precursor related cluster 239 2e-62

Implicit cross-link from Swiss-Prot/TrEMBL to UniRef:

Other	
RZPD-ProtExp	C0102; -.
SOURCE	ALPP; Homo sapiens.
ProtoNet	P05187.
UniRef	View cluster of proteins with at least 50% / 90% / 100% identity.

UniRef90 Entry

UniRef90_P01588	Erythropoietin precursor related cluster

Member Sequence

UniProtKB

UniParc

UniProt/UniParc ID	UniProt Accessions	UniRef100 ID	Protein Name	Species Name	Taxon ID	Length
EPO_HUMAN	P01588	UniRef100_P01588	Erythropoietin precursor	Homo sapiens (Human)	9606	193
Q2M2L6_HUMAN	Q2M2L6	UniRef100_P01588	Erythropoietin,	Homo sapiens (Human)	9606	193
EPO_MACFA	P07865	UniRef100_P07865	Erythropoietin precursor	Macaca fascicularis (Crab eating macaque) (Cynomolgus monkey)	9541	192
EPO_MACMU	Q28513	UniRef100_Q28513	Erythropoietin precursor	Macaca mulatta (Rhesus macaque)	9544	192
UPI0000110A62	UPI0000110A62	UniRef100_UPI0000110A62	PROTEIN (ERYTHROPOIETIN)	Homo sapiens	9606	166
UPI00001112A2	UPI00001112A2	UniRef100_UPI00001112A2	ERYTHROPOIETIN	Homo sapiens	9606	166
Q8HZ86_9PRIM	Q8HZ86	UniRef100_Q8HZ86	Erythropoietin	Macaca sp	9549	133
Q8HZ88_9PRIM	Q8HZ88	UniRef100_Q8HZ88	Erythropoietin	Gorilla gorilla (gorilla)	9593	133
Q8HZ89_PANTR	Q8HZ89	UniRef100_Q8HZ89	Erythropoietin	Pan troglodytes (Chimpanzee)	9598	133
Q8HZ87_PONPY	Q8HZ87	UniRef100_Q8HZ87	Erythropoietin	Pongo pygmaeus (Orangutan)	9600	131

Representative Sequence (AC=P01588)

MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAENITTGCAEHC
SLNENITVPDTKVNFYAWKRMEVGQQAVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQL
HVDKAVSGLRSLTTLLRALGAQKEAISPPDAASAAPLRTITADTFRKLFRVYSNFLRGKL
KLYTGEACRTGDR

UniParc – the UniProt Archive

- Sequences and cross-references
- A comprehensive collection of the raw protein sequences in public databases (including those not submitted to the DNA databases):

  Swiss-Prot, TrEMBL, PIR, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, Patent Offices.
- UniParc allows to track sequence versions

Use with extreme caution: also contains pseudogenes, incorrect CDS predictions, etc…and highly redundant !

UniParc allows to keep track of a protein sequence and of its integration in various databases

Viewers: XML | ExPASy | SRS | PIR

UPI		UPI0000033477					
Sequence		MGVHECPAWL WLLLSLLSLP LGLPVLGAPP RLICDSRVLE RYLLEAKEAE NITTGCAEHC SLNENITVPD TKVNFYAWKR MEVGQQAVEV WQGLALLSEA VLRGQALLVN SSQPWEPLQL HVDKAVSGLR SLTTLLRALG AQKEAISPPD AASAAPLRTI TADTFRKLFR VYSNFLRGKL KLYTGEACRT GDR					
	Length	193					
	CRC64	C91F0E4C26A52033					
References	DataBase	Accession	Version	Active	Created	Last Update	Deleted
	EMBL	AAA52400.1	1	Y	12-MAR-2003	21-MAR-2006	-
	EMBL	AAC78791.1	1	Y	12-MAR-2003	21-MAR-2006	-
	EMBL	AAF23132.1	1	Y	12-MAR-2003	21-MAR-2006	-
	EMBL	AAF23134.1	1	Y	12-MAR-2003	21-MAR-2006	-
	EMBL	AAH93628.1	1	Y	20-APR-2005	21-MAR-2006	-
	EMBL	AAI11938.1	1	Y	23-JAN-2006	21-MAR-2006	-
	EMBL	AAP22357.1	1	Y	16-JUN-2003	21-MAR-2006	-
	EMBL	CAC09044.1	1	Y	12-MAR-2003	21-MAR-2006	-
	EPO	AX025443.1	1	Y	26-MAR-2003	14-JUN-2006	-
	EPO	AX046871.1	1	Y	26-MAR-2003	14-JUN-2006	-
	EPO	AX320725.1	1	Y	26-MAR-2003	14-JUN-2006	-
	EPO	AX591241.1	1	Y	26-MAR-2003	14-JUN-2006	-
	EPO	AX644923.1	1	Y	26-MAR-2003	14-JUN-2006	-
	Ensembl(Human)	ENSP00000252723	1	Y	01-APR-2003	03-APR-2006	-
	IPI	IPI00003100.1	1	N	14-MAR-2003	-	10-APR-2003
	IPI	IPI00307226.1	1	N	13-JUN-2003	-	03-OCT-2003
	IPI	IPI00307226.3	3	Y	19-NOV-2003	13-JUN-2006	-
	JPO	BD617478	1	Y	06-JUL-2004	19-MAY-2006	-

Patent data

UniParc entry UPI0000033477 part 2

RefSeq	NP_000790.2	2	Y	07-APR-2005	08-MAY-2006	-
RemTrEMBL	CAC09044	1	N	28-MAR-2003	-	17-NOV-2003
Swiss-Prot	P01588.1	1	Y	01-NOV-1988	27-JUN-2006	-
TROME(Human)	NT_007933_293_0	2	N	30-DEC-2003	-	16-APR-2004
TROME(Human)	NT_007933_293_1	2	N	30-DEC-2003	-	16-APR-2004
TROME(Human)	NT_007933_299_101	1	N	25-FEB-2005	-	25-FEB-2005
TROME(Human)	NT_007933_299_34	1	N	25-FEB-2005	-	25-FEB-2005
TROME(Human)	NT_007933_301_32	1	N	23-AUG-2004	-	28-AUG-2004
TROME(Human)	NT_007933_301_87	1	N	23-AUG-2004	-	28-AUG-2004
TROME(Human)	NT_007933_303_101	1	N	28-NOV-2004	-	20-FEB-2005
TROME(Human)	NT_007933_303_35	1	N	28-NOV-2004	-	20-FEB-2005
TROME(Human)	NT_007933_315_0	3	N	13-AUG-2004	-	13-AUG-2004
TROME(Human)	NT_007933_315_1	3	N	13-AUG-2004	-	13-AUG-2004
TROME(Human)	NT_007933_326_0	1	N	11-NOV-2003	-	02-DEC-2003
TROME(Human)	NT_007933_326_1	1	N	11-NOV-2003	-	02-DEC-2003
TROME(Human)	NT_007933_366_0	6	Y	03-AUG-2005	04-AUG-2005	
TROME(Human)	NT_007933_366_1	6	Y	03-AUG-2005	04-AUG-2005	
TROME(Human)	NT_079595_293_0	1	N	30-DEC-2003	-	16-APR-2004
TROME(Human)	NT_079595_293_1	1	N	30-DEC-2003	-	16-APR-2004
TROME(Human)	NT_079595_322_0	2	N	23-AUG-2004	-	28-AUG-2004
TROME(Human)	NT_079595_322_1	2	N	23-AUG-2004	-	28-AUG-2004
TROME(Human)	NT_079595_323_0	1	N	13-AUG-2004	-	13-AUG-2004
TROME(Human)	NT_079595_323_1	1	N	13-AUG-2004	-	13-AUG-2004
TrEMBL	Q2M2L6.1	1	Y	21-FEB-2006	27-JUN-2006	-
TrEMBL	Q549U2.1	1	N	24-MAY-2005	-	11-OCT-2005
TrEMBLnew	AAP22357	1	N	03-MAY-2003	-	11-JUN-2004
UPO	AAA00944.1	1	Y	26-MAR-2003	19-MAY-2006	
UPO	AAA54969.1	1	Y	26-MAR-2003	19-MAY-2006	-
UPO	AAA54970.1	1	Y	26-MAR-2003	19-MAY-2006	-
UPO	AAA55466.1	1	Y	26-MAR-2003	19-MAY-2006	-
UPO	AAA56085.1	1	Y	26-MAR-2003	19-MAY-2006	-

TrEMBL entry probably to be merged into Swiss-Prot

TrEMBL entry was merged into Swiss-Prot

« Non-Redundancy » definitions

- UniParc
  - One UniParc entry for all 100% identical sequences (from many different databases)
- UniRef
  - One UniRef100 entry for all 100% identical sequences (including subfragments) from the UniProt Knowledgebase
- Swiss-Prot
  - One Swiss-Prot entry per gene/protein, including fragments, variations, splice variants

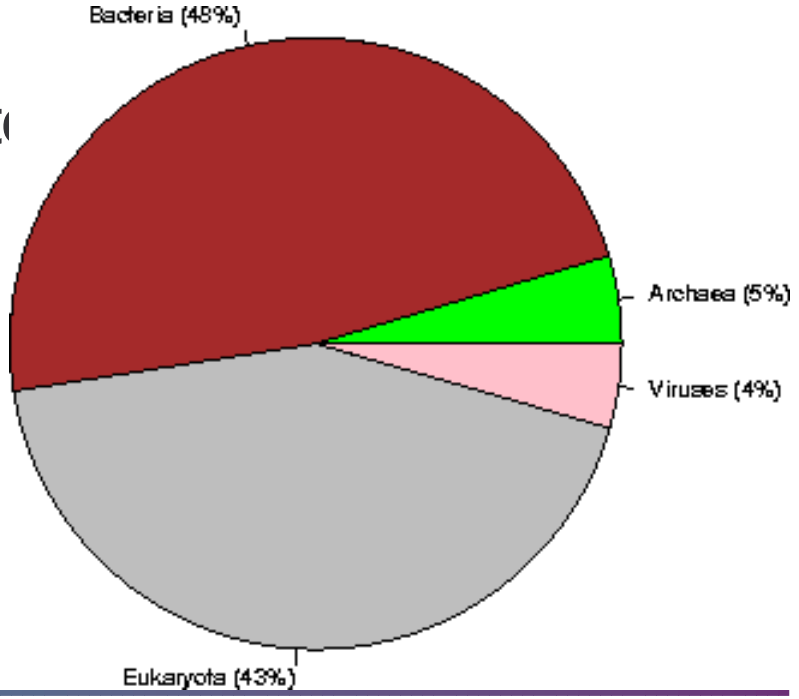
How many species are represented in Swiss-Prot ?

http://www.expasy.org/sprot/relnotes/relstat.html

- Almost 10'000 different species;
- 20 species represent about 35% of all sequences in the database;
- >7'000 species are only represent

by one to three sequences.

In most cases these are sequences which were done in the context of a phylogenetic study.



Swiss-Prot annotation priorities

The main annotation programs:

- HAMAP (High quality Automated and Manual Annotation of microbial Proteomes; bacteria, archaea, chloroplasts, mitochondria);
- HPI (Human Proteomics Initiative);
- PPAP (Plant Proteome Annotation Project);
- FPAP (Fungal Proteome Annotation Project);
- Viral Proteins;
- Tox-Prot (Toxin Annotation Project);
- ENZYMES (proteins with EC numbers);
- PTMs
- 3d-structure

Model organisms

- Organisms for which we want to have a more in-depth coverage;
- Completeness, links with specialized databases, specific documents;
- Examples: E.coli, B.subtilis, Human, Mouse, Fly, C.elegans, Yeast, S.pombe, A.thaliana.

Human proteomics initiative (HPI) status report

UniProtKB/Swiss-Prot Release 50.1 of 13-Jun-2006: 223100 entries (Release statistics)

Total number of annotated mammalian sequences in UniProtKB/Swiss-Prot: 43070

Total number of annotated human sequences in UniProtKB/Swiss-Prot: 14094

		max per entry	average per entry	number of entries
Number of isoforms due to alternative splicing, initiation or promoter usage:	**7707**	32	0.55	4129 (29.30%)
Number of variants (disease mutations and polymorphisms):	**26439**	252	1.88	4390 (31.15%)
Number of annotated post-translational modifications (experimentally proven or potential):	**38255**	212	2.71	7120 (50.52%)
Number of references to published articles:	**65464** (35597 distinct references)	152	4.64	13845 (98.23%)
Number of comment blocks:	**77984**	30	5.53	13774 (97.73%)
Number of feature lines:	**349393**	761	24.79	14094 (100.00%)
Number of cross-referenced EMBL protein_ids:	**68538** (68472 distinct protein_ids)	539	4.86	13927 (98.82%)
Number of cross-references to InterPro:	**37315**	21	2.65	12744 (90.42%)
Number of cross-references to PDB (3D-structure):	**8454**	185	0.60	2203 (15.63%)
Number of cross-references to MIM:	**11944** (11281 distinct MIM entries)	13	0.85	9753 (69.20%)
Number of cross-references to HGNC:	**13497** (13369 distinct HGNC entries)	13	0.96	13439 (95.35%)

Swiss-Prot documents

http://www.expasy.org/sprot/sp-docu.html
ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/complete/docs

About 90 different documents:

- User manual, release notes;
- Indices (authors, citations, keywords, etc.);
- Lists per species or per chromosome;
- Nomenclature documents;
- User-maintained specialized documents

UniProtKB/Swiss-Prot Documentation

Most of the documents listed in this page can also be downloaded by ftp (versions without hypertext links).

[General documents] [Nomenclature documents] [Species-specific documents] [Other documents]

General documents

- User manual for the UniProt Knowledgebase (UniProtKB/Swiss-Prot+UniProtKB/TrEMBL)
- Release notes for the current **major** release / Statistics for the current **biweekly** release: UniProtKB/Swiss-Prot + UniProtKB/TrEMBL
- Recent format changes: flat file format / XML format
- Forthcoming format changes: flat file format / XML format

- UniProtKB/Swiss-Prot annotation: how biochemical information is assigned to sequence entries
- List of on-line experts
- List of abbreviations for journals cited
- List of keywords and definition of their usage
- List of organism identification codes
- List of tissues
- List of strains
- List of plasmids
- List of post-translational modifications
- List of databases cross-referenced in UniProtKB/Swiss-Prot
- Index of CC Pathway lines in UniProtKB/Swiss-Prot [Browse]
- Index of CC SIMILARITY lines in UniProtKB/Swiss-Prot [Browse]

http://www.expasy.org/sprot/sp-docu.html

Nomenclature documents

- Protein naming guidelines new
- List of nomenclature related references for proteins
- Nomenclature of extracellular domains
- Blood group antigens proteins

New document: Protein naming guidelines

A **"recommended name" (RN)** should be unique and attributed to all orthologs. One reason for this is that it should be possible to propagate a protein name to all orthologous proteins, from various organisms. This is why, ideally, the protein name should not contain a specific characteristic of the protein, and in particular it should not reflect the function or role of the protein, nor its subcellular location, its domain structure, its tissue specificity, its molecular weight or its species of origin.

Therefore a RN should :

- not contain information about the molecular weight of the protein. e.g. "Unicornase subunit A" is preferred to "Unicornase 52 kDa subunit."
- not be based on the name of a disease. e.g. "Bloom syndrome protein" is not suitable.
- not be based on tissue specificity. e.g. "Testis-specific protein ..." is not suitable.
- not be based on the species name. e.g. "Yeast Ku70 protein" is not suitable.
- not be based on the gene induction. e.g. "Androgen-induced protein 1" is not suitable.

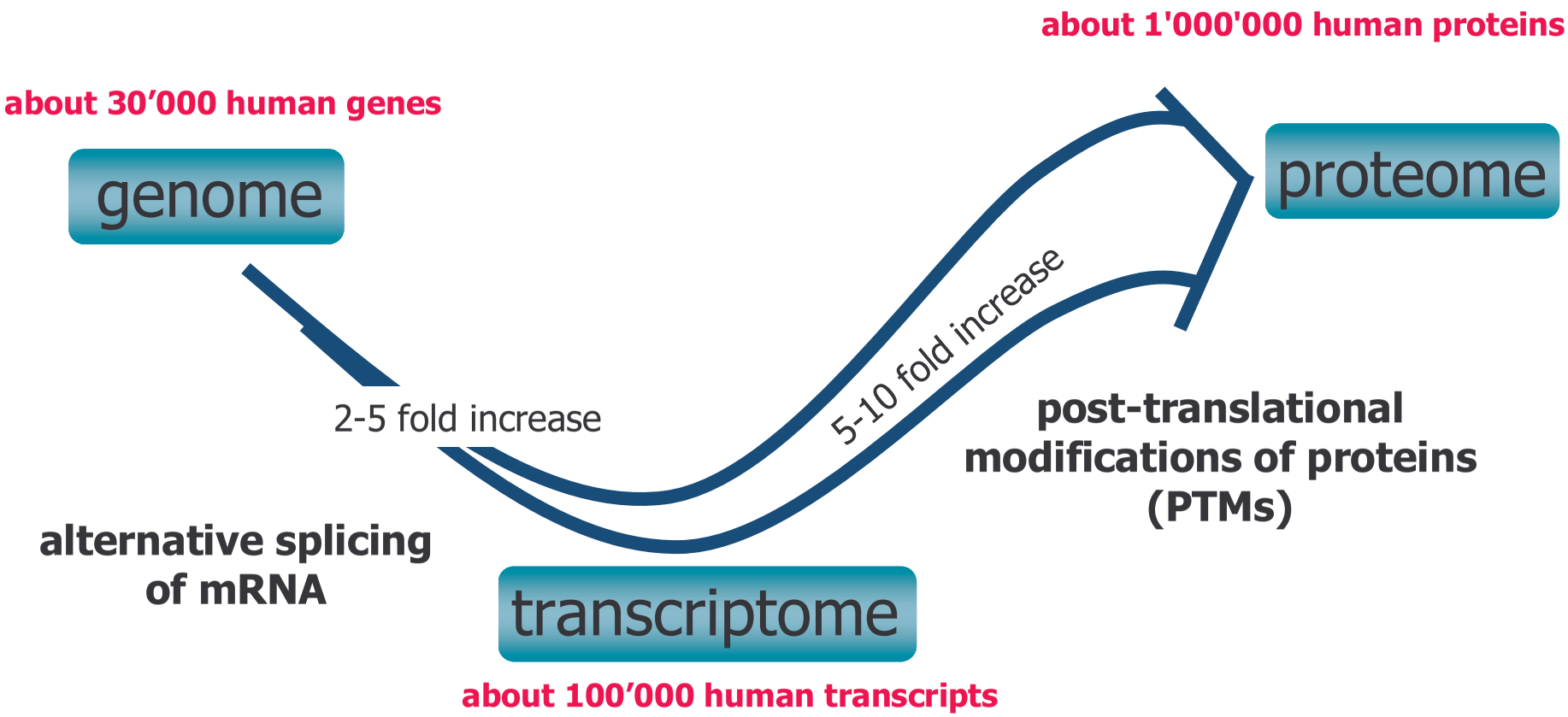
The most optimal RN is a word that ends with "in" and which can be easily pronounced in English. e.g. "Zyxin", "Insulin", "Hemoglobin", "Caveolin", "Desmoglein", "Secretin", etc.

UniProtKB/Swiss-Prot annotation and protein diversity

1. Alternative splicing.
2. Variants.
3. Post-translational modifications.

Boeckmann B. et al. *Protein variety and functional diversity: Swiss-Prot annotation in its biological context*
Comptes Rendus Biologies 328:882-99(2005).

From genome to proteome



Estimation of % alternatively spliced human genes

Mironov et al., 1999	> 35 %
Croft et al., 2000	> 22 %
HGC, 2001	> 59 %
Kan et al., 2001	> 55 %
Modrek et al., 2001	> 42 %
Brett et al., 2002	> 50 %

20% of all human Swiss-Prot entries have annotation about alternative splicing

More than half of the changes brought about by splicing are simple: missing/inserted N-, C- or internal regions; the rest can be very complex.

Swiss-Prot & alternative splicing

- Annotation at two levels: comments and features;
- The sequence record generally represents the longest isoform;
- Stable isoform identifiers and feature identifiers have been introduced to facilitate access to splice isoforms
- Many tools on ExPASy have been adapted to distinguish the various splice isoforms.

Keywords

Keywords	2A5G_HUMAN, Q13362
Alternative splicing; Hydrolase; Multigene family; Nuclear protein; Phosphorylation; Protein phosphatase.	

P73_HUMAN (**O15350**) references to alternative isoforms

References

[1] NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS ALPHA AND BETA).
TISSUE=Colon;
DOI=10.1016/S0092-8674(00)80540-1; PubMed=9288759 [NCBI, ExPASy, EBI, Israel, Japan]
Kaghad M., Bonnet H., Yang A., Creancier L., Biscan J.-C., Valent A., Minty A., Chalon P., Lelias J.-M., Dumont X., Ferrara P., McKeon F., Caput D.;
"Monoallelically expressed gene related to p53 at 1p36, a region frequently deleted in neuroblastoma and other human cancers.";
Cell 90:809-819(1997).

[2] NUCLEOTIDE SEQUENCE [GENOMIC DNA] (ISOFORM ALPHA).
DOI=10.1006/geno.1998.5387; PubMed=9721206 [NCBI, ExPASy, EBI, Israel, Japan]
Mai M., Huang H., Reed C., Qian C., Smith J.S., Alderete B., Jenkins R., Smith D.I., Liu W.;
"Genomic organization and mutation analysis of p73 in oligodendrogliomas with chromosome 1 p-arm deletions.";
Genomics 51:359-363(1998).

[3] NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS GAMMA AND DELTA).
TISSUE=Neuroblastoma;
DOI=10.1084/jem.188.9.1763; PubMed=9802988 [NCBI, ExPASy, EBI, Israel, Japan]
De Laurenzi V., Costanzo A., Barcaroli D., Terrinoni A., Falco M., Annicchiarico-Petruzzelli M., Levrero M., Melino G.;
"Two new p73 splice variants, gamma and delta, with different transcriptional activity.";
J. Exp. Med. 188:1763-1768(1998).

[4] NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS EPSILON AND ZETA).
TISSUE=Hepatoma, Lymphocyte, Mammary cancer, and Skin;
DOI=10.1038/sj.cdd.4400521; PubMed=10381648 [NCBI, ExPASy, EBI, Israel, Japan]
De Laurenzi V., Catani M.V., Terrinoni A., Corazzari M., Melino G., Costanzo A., Levrero M., Knight R.A.;
"Additional complexity in p73: induction by mitogens in lymphoid cells and identification of two new splicing variants epsilon and zeta.";
Cell Death Differ. 6:389-390(1999).

[5] NUCLEOTIDE SEQUENCE [GENOMIC DNA] (ISOFORM ALPHA).
DOI=10.1038/sj.onc.1202677; PubMed=10362363 [NCBI, ExPASy, EBI, Israel, Japan]
Yoshikawa H., Nagashima M., Khan M.A., McMenamin M.G., Hagiwara K., Harris C.C.;
"Mutational analysis of p73 and p53 in human cancer cell lines.";
Oncogene 18:3415-3421(1999).

[6] NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS DN-ALPHA; DN-BETA AND DN-GAMMA).
DOI=10.1038/sj.cdd.4400962; PubMed=11753569 [NCBI, ExPASy, EBI, Israel, Japan]
Grob T.J., Novak U., Maisse C., Barcaroli D., Luthi A.U., Pirnia F., Hugli B., Graber H.U., De Laurenzi V., Fey M.F., Melino G., Tobler A.;
"Human DeltaNp73 regulates a dominant negative feedback loop for TAp73 and p53.";
Cell Death Differ. 8:1213-1223(2001).

P73_HUMAN (O15350) Comment topic « Alternative products »

- *ALTERNATIVE PRODUCTS*: 9 named isoforms [FASTA] produced by alternative splicing.

Name	**Alpha**
Isoform ID	O15350-1
This is the isoform sequence displayed in this entry.	

Name	**Beta**
Isoform ID	O15350-2
Features which should be applied to build the isoform sequence: VSP_006539.	

Name	**Gamma**
Isoform ID	O15350-3
Note: The splicing of exon 11 results in a frameshift from the original reading frame.	
Features which should be applied to build the isoform sequence: VSP_006540, VSP_006541.	

Name	**Delta**
Isoform ID	O15350-4
Features which should be applied to build the isoform sequence: VSP_006542, VSP_006543.	

Name	**Epsilon**
Isoform ID	O15350-5
Note: The splicing of exon 11 results in a frameshift from the original reading frame. The splicing of exon 13 reverts the reading frame to the sequence of isoform Alpha.	
Features which should be applied to build the isoform sequence: VSP_006544, VSP_006545.	

Name	**Zeta**
Isoform ID	O15350-6
Features which should be applied to build the isoform sequence: VSP_006546.	

Name	**dN-Alpha**
Isoform ID	O15350-8
Features which should be applied to build the isoform sequence: VSP_014368.	

Name	**dN-Beta**
Isoform ID	O15350-9

Unique and stable isoform IDs

Feature IDs

Alternative splicing in feature tables – and links to reconstituted sequences

VAR_SEQ	1	62	MAQSTATSPDGGTTFEHLWSSLEPDSTYFDLPQSSRGNNE VVGGTDSSMDVFHLEGMTTSVM -> MLYVGDPARHLAT (in isoform dN-Alpha, isoform dN-Beta and isoform dN-Gamma).	VSP_014368
VAR_SEQ	400	495	Missing (in isoform Zeta).	VSP_006546
VAR_SEQ	400	476	SHLQPPSYGPVLSPMNKVHGGMNKLPSVNQLVGQPPPHSS AATPNLGPVGPGMLNNHGHAVPANGEMSSSHSAQSMV -> PRDAQQPWPRSASQQRRDEQQPQRPVHGLGVPLHSATPLP RRPQPRQFFNRIQVSKLHRVFHLPRVTEHLPPAEPDH (in isoform Gamma and isoform dN-Gamma).	VSP_006540
VAR_SEQ	400	445	SHLQPPSYGPVLSPMNKVHGGMNKLPSVNQLVGQPPPHSS AATPNL -> PRDAQQPWPRSASQQRRDEQQPQRPVHGLGVPLHSATPLP RRPQPR (in isoform Epsilon).	VSP_006544
VAR_SEQ	400	403	SHLQ -> TWGP (in isoform Delta).	VSP_006542
VAR_SEQ	404	636	Missing (in isoform Delta).	VSP_006543
VAR_SEQ	446	526	Missing (in isoform Epsilon).	VSP_006545
VAR_SEQ	477	636	Missing (in isoform Gamma and isoform dN-Gamma).	VSP_006541
VAR_SEQ	495	636	SFLTGLGCPNCIEYFTSQGLQSIYHLQNLTIEDLGALKIP EQYRMTIWRGLQDLKQGHDYSTAQQLLRSSNAATISIGGS GELQRQRVMEAVHFRVRHTITIPNRGGPGGGPDEWADFGF DLPDCKARKQPIKEEFTEAEIH -> RTWGP (in isoform Beta and isoform dN-Beta).	VSP_006539

Isoform description	
Name	**Gamma**
Isoform ID	O15350-3
Note: The splicing of exon 11 results in a frameshift from the original reading frame.	
Features which should be applied to build the isoform sequence	VSP_006540, VSP_006541

Isoform description

Sequence information

Length: **476 AA**

```
        10         20         30         40         50         60
MAQSTATSPD GGTTFEHLWS SLEPDSTYFD LPQSSRGNNE VVGGTDSSMD VFHLEGMTTS

        70         80         90        100        110        120
VMAQFNLLSS TMDQMSSRAA SASPYTPEHA ASVPTHSPYA QPSSTFDTMS PAPVIPSNTD

       130        140        150        160        170        180
YPGPHHFEVT FQQSSTAKSA TWTYSPLLKK LYCQIAKTCP IQIKVSTPPP PGTAIRAMPV

       190        200        210        220        230        240
YKKAEHVTDV VKRCPNHELG RDFNEGQSAP ASHLIRVEGN NLSQYVDDPV TGRQSVVVPY

       250        260        270        280        290        300
EPPQVGTEFT TILYNFMCNS SCVGGMNRRP ILIIITLEMR DGQVLGRRSF EGRICACPGR

       310        320        330        340        350        360
DRKADEDHYR EQQALNESSA KNGAASKRAF KQSPPAVPAL GAGVKKRRHG DEDTYYLQVR

       370        380        390        400        410        420
GRENFEILMK LKESLELMEL VPQPLVDSYR QQQQLLQRPP RDAQQPWPRS ASQQRRDEQQ

       430        440        450        460        470
PQRPVHGLGV PLHSATPLPR RPQPRQFFNR IGVSKLHRVF HLPRVTEHLP PAEPDH
```

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

Reconstituted sequence

Direct links to analysis tools

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

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

Splice isoform list of Swiss-Prot entry: O15350

Send selected sequences to Clustal W (multiple alignment) Submit Query Select all

- Clustal W (multiple alignment)
- T-COFFEE (multiple alignment)
- Reduce redundancy
- PRATT (find conserved patterns)
- Retrieve entries (Swiss-Prot format)
- Retrieve sequences (FASTA format)
- Retrieve matching entries accession codes

>sp|O15350|P73_HUMAN cription factor) (p53-related protein) -
MAQSTATSPDGGTTFEHLWSSLE TTS
VMAQFNLLSSTMDQMSSRAASAS NTD
YPGPHHFEVTFQQSSTAKSATWI MPV
YKKAEHVTDVVKRCPNHELGRDF VPY
EPPQVGTEFTTILYNFMCNSSCVGGMNRRPILIIITLEMRDGQVLGRRSFEGRICACPGR
DRKADEDHYREQQALNESSAKNGAASKRAFKQSPPAVPALGAGVKKRRHGDEDTYYLQVR
GRENFEILMKLKESLELMELVPQPLVDSYRQQQQLLQRPSHLQPPSYGPVLSPMNKVHGG
MNKLPSVNQLVGQPPPHSSAATPNLGPVGPGMLNNHGHAVPANGEMSSSHSAQSMVSGSH
CTPPPPYHADPSLVSFLTGLGCPNCIEYFTSQGLQSIYHLQNLTIEDLGALKIPEQYRMT
IWRGLQDLKQGHDYSTAQQLLRSSNAATISIGGSGELQRQRVMEAVHFRVRHTITIPNRG
GPGGGPDEWADFGFDLPDCKARKQPIKEEFTEAEIH

>sp_vs|O15350-2|O15350 Splice isoform Beta of O15350
MAQSTATSPDGGTTFEHLWSSLEPDSTYFDLPQSSRGNNEVVGGTDSSMDVFHLEGMTTS
VMAQFNLLSSTMDQMSSRAASASPYTPEHAASVPTHSPYAQPSSTFDTMSPAPVIPSNTD
YPGPHHFEVTFQQSSTAKSATWTYSPLLKKLYCQIAKTCPIQIKVSTPPPPGTAIRAMPV
YKKAEHVTDVVKRCPNHELGRDFNEGQSAPASHLIRVEGNNLSQYVDDPVTGRQSVVVPY
EPPQVGTEFTTILYNFMCNSSCVGGMNRRPILIIITLEMRDGQVLGRRSFEGRICACPGR
DRKADEDHYREQQALNESSAKNGAASKRAFKQSPPAVPALGAGVKKRRHGDEDTYYLQVR
GRENFEILMKLKESLELMELVPQPLVDSYRQQQQLLQRPSHLQPPSYGPVLSPMNKVHGG
MNKLPSVNQLVGQPPPHSSAATPNLGPVGPGMLNNHGHAVPANGEMSSSHSAQSMVSGSH
CTPPPPYHADPSLVRTWGP

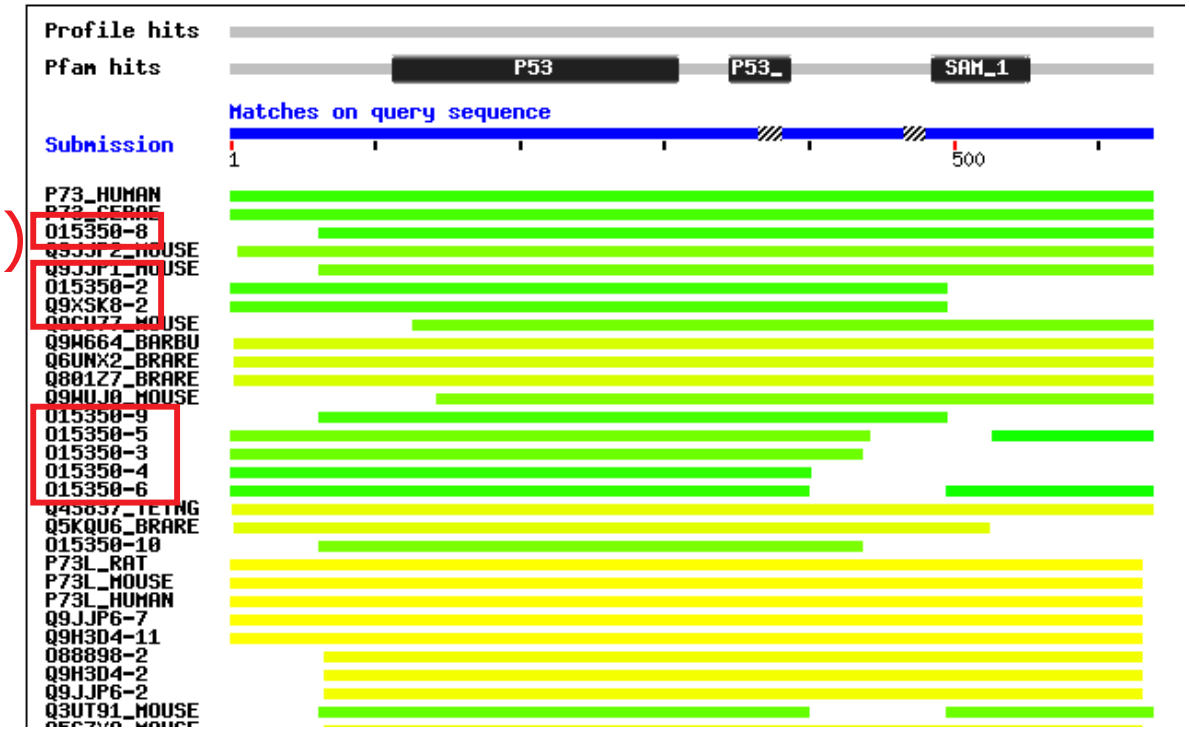
>sp_vs|O15350-3|O15350 Splice isoform Gamma of O15350
MAQSTATSPDGGTTFEHLWSSLEPDSTYFDLPQSSRGNNEVVGGTDSSMDVFHLEGMTTS
VMAQFNLLSSTMDQMSSRAASASPYTPEHAASVPTHSPYAQPSSTFDTMSPAPVIPSNTD
YPGPHHFEVTFQQSSTAKSATWTYSPLLKKLYCQIAKTCPIQIKVSTPPPPGTAIRAMPV
YKKAEHVTDVVKRCPNHELGRDFNEGQSAPASHLIRVEGNNLSQYVDDPVTGRQSVVVPY
EPPQVGTEFTTILYNFMCNSSCVGGMNRRPILIIITLEMRDGQVLGRRSFEGRICACPGR

Submit splice isoform sequences to further analysis, e.g. multiple aligment

ExPASy tools and alternative isoforms

- Aldente (protein identification by peptide mass fingerprinting)
- ScanProsite (pattern searching)
- Blast

O15350 (P73_HUMAN) and its alternative isoform sequences are found separately



UniProtKB/Swiss-Prot annotation and protein diversity

1. Alternative splicing.
2. Variants.
3. Post-translational modifications.

Boeckmann B. et al. *Protein variety and functional diversity: Swiss-Prot annotation in its biological context*
Comptes Rendus Biologies 328:882-99(2005).

Polymorphisms

- Non synonymous c-**SNPs** (coding **S**ingle **N**ucleotide **P**olymorphism**s**) (**S**ingle **A**mino-acid **P**olymorphism**s** or **SAPs**);

- Mutations causing profound changes, such as a frameshift or a STOP codon, are not annotated (deleterious effect on function is obvious);

- Annotation can be found in 4 sections:
  - references;
  - comments (disease or polymorphism);
  - keywords (disease mutation or polymorphism);
  - features (FT VARIANT).

Variants & References

Polymorphism

Disease variant: Li-Fraumeni syndrome

P53_HUMAN, P04637

[27] VARIANT ARG-72.
MEDLINE=91153807; PubMed=1999338; [NCBI, ExPASy, EBI, Israel, Japan]
Olschwang S., Laurent-Puig P., Vassal A., Salmon R.-J., Thomas G.;
"Characterization of a frequent polymorphism in the coding sequence of the Tp53 gene in colonic cancer patients and a control population.";
Hum. Genet. 86:369-370(1991).

[28] VARIANT LFS THR-133.
MEDLINE=92034774; PubMed=1933902; [NCBI, ExPASy, EBI, Israel, Japan]
Law J.C., Strong L.C., Chidambaram A., Ferrell R.E.;
"A germ line mutation in exon 5 of the p53 gene in an extended cancer family.";
Cancer Res. 51:6385-6387(1991).

[29] VARIANTS LFS CYS-245; TRP-248; PRO-252 AND LYS-258.
MEDLINE=91057657; PubMed=1978757; [NCBI, ExPASy, EBI, Israel, Japan]
Malkin D., Li F.P., Strong L.C., Fraumeni J.F. Jr., Nelson C.E., Kim D.H., Kassel J., Gryka M.A., Bischoff F.Z., Tainsky M.A., Friend S.H.;
"Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms.";
Science 250:1233-1238(1990).

[30] VARIANT LFS ASP-245.
MEDLINE=91080929; PubMed=2259385; [NCBI, ExPASy, EBI, Israel, Japan]
Srivastava S., Zou Z., Pirollo K., Blattner W., Chang E.H.;
"Germ-line transmission of a mutated p53 gene in a cancer-prone family with Li-Fraumeni syndrome.";
Nature 348:747-749(1990).

[31] VARIANT LFS LEU-272.
MEDLINE=92147883; PubMed=1737852; [NCBI, ExPASy, EBI, Israel, Japan]
Felix C.A., Nau M.M., Takahashi T., Mitsudomi T., Chiba I., Poplack D.G., Reaman G.H., Cole D.E., Letterio J.J., Whang-Peng J., Knutsen T., Minna J.D.;
"Hereditary and acquired p53 gene mutations in childhood acute lymphoblastic leukemia.";
J. Clin. Invest. 89:640-647(1992).

[32] VARIANTS LFS HIS-273 AND VAL-325.
MEDLINE=92228023; PubMed=1565144; [NCBI, ExPASy, EBI, Israel, Japan]
Malkin D., Jolly K.W., Barbier N., Look A.T., Friend S.H., Gebhardt M.C., Andersen T.I., Boerresen A.-L., Li F.P., Garber J., Strong L.C.;
"Germline mutations of the p53 tumor-suppressor gene in children and young adults with second malignant neoplasms.";

Variants & Comments

Comments P53_HUMAN, P04637

- ***FUNCTION***: Acts as a tumor suppressor in many tumor types; induces growth arrest or apoptosis depending on the physiological circumstances and cell type. Involved in cell cycle regulation as a trans-activator that acts to negatively regulate cell division by controlling a set of genes required for this process. One of the activated genes is an inhibitor of cyclin-dependent kinases. Apoptosis induction seems to be mediated either by stimulation of BAX and FAS antigen expression, or by repression of Bcl-2 expression.
- ***DISEASE***: TP53 is found in increased amounts in a wide variety of transformed cells. TP53 is frequently mutated or inactivated in about 60% of cancers.
- ***DISEASE***: Defects in TP53 are involved in esophageal squamous cell carcinoma (ESCC) [MIM:133239]. ESCC is a tumor of the esophagus.
- ***DISEASE***: Defects in TP53 are a cause of Li-Fraumeni syndrome (LFS) [MIM:151623]. LFS is an autosomal dominant familial cancer syndrome that in its classic form is defined by the existence of both a proband with a sarcoma and two other first-degree relatives with a cancer by age 45 years. In these families the affected relatives develop a diverse set of malignancies at unusually early ages. The spectrum of cancers in LFS includes breast carcinomas, soft-tissue sarcomas, brain tumors, osteosarcoma, leukemia and adreno-cortical carcinoma. Other possible component tumors of LFS are melanoma, gonadal cell tumors and carcinomas of the lung, pancreas and prostate.
- ***DISEASE***: Defects in TP53 are found in Barrett metaplasia; also known as Barrett esophagus. It is a condition in which the normally stratified squamous epithelium of the lower esophagus is replaced by a metaplastic columnar epithelium. The condition develops as a complication in approximately 10% of patients with chronic gastroesophageal reflux disease and predisposes to the development of esophageal adenocarcinoma.
- ***DISEASE***: Defects in TP53 are involved in head and neck squamous cell carcinomas (HNSCC) [MIM:275355].
- ***DISEASE***: Defects in TP53 are involved in oral squamous cell carcinoma (OSCC). Cigarette smoke is a prime mutagenic agent in cancer of the aerodigestive tract.
- ***DISEASE***: Defects in TP53 are a cause of lung cancer [MIM:211980].
- ***DISEASE***: Defects in TP53 are a cause of choroid plexus papilloma [MIM:260500]. Choroid plexus papilloma is a slow-growing benign tumor of the choroid plexus that often invades the leptomeninges. In children it is usually in a lateral ventricle but in adults it is more often in the fourth ventricle. Hydrocephalus is common, either from obstruction or from tumor secretion of cerebrospinal fluid. If it undergoes malignant transformation it is called a choroid plexus carcinoma. Primary choroid plexus tumors are rare and usually occur in early childhood.
- ***SIMILARITY***: Belongs to the p53 family.
- ***WEB RESOURCE***: NAME=IARC TP53 mutation database; NOTE=Somatic and germline TP53 mutations in human cancers;

#151623

Links

LI-FRAUMENI SYNDROME; LFS

Alternative titles; symbols

SARCOMA FAMILY SYNDROME OF LI AND FRAUMENI
SBLA SYNDROME
LI-FRAUMENI SYNDROME-VARIANT, INCLUDED
LFS-VARIANT, INCLUDED

Gene map locus 22q12.1, 17p13.1, 9p21

TEXT

A number sign (#) is used with this entry because mutations in the TP53 gene (191170) have been found in this disorder. Additionally, mutations in the CHK2 gene (604373) have been found to cause Li-Fraumeni syndrome.

In reviewing medical records and death certificates of 648 childhood rhabdomyosarcoma patients, Li and Fraumeni (1969) identified 4 families in which sibs or cousins had a childhood sarcoma. These 4 families also had striking histories of breast cancer and other neoplasms, suggesting a new familial cancer syndrome of diverse tumors. Subsequent prospective studies confirmed the high risk in family members of the tumor types that comprise LFS (Li and Fraumeni, 1982). Studies in other geographic and ethnic groups by Birch et al. (1984, 1990) corroborated the syndrome. The spectrum of cancers in the syndrome was shown to include, in addition to breast cancer and soft tissue sarcomas, brain tumors, osteosarcoma, leukemia, and adrenocortical carcinoma.

Fraumeni et al. (1975) described a kindred in which in 1 sibship of 9 adults, 4 died of lymphocytic or histiocytic lymphomas and one, a male, of Waldenstrom macroglobulinemia complicated by adenocarcinoma of the lung. In the next generation, 1 person died of Hodgkin disease; 4 of 9 healthy persons had impaired lymphocyte transformation with phytohemagglutinin and 3 of these had polyclonal elevation of IgM. Subsequent to the studies, adenocarcinoma of the lung developed in one of those with an immune defect, a woman, and her 3-year-old grandson developed lymphocytic leukemia. This is the first suggestion of a genetic or immunologic basis of lung adenocarcinoma. Pearson et al. (1982) reported 2 families resembling that reported by Li and Fraumeni (1969). In 1, the mother had breast cancer and 3 of her 4 children had adrenocortical carcinoma, medulloblastoma, and rhabdomyosarcoma; in the other, the mother had breast cancer and 2 of her 3 children had adrenocortical carcinoma and

Variants & Comments

If the variant is NOT associated with a disease state…

HMT_HUMAN, P50135
POLYMORPHISM: Variant Ile-105 has a reduced activity and seems to be linked with a predisposition to asthma.

MYOC_MOUSE, O70624
POLYMORPHISM: Variant Ala-164 is found in strain BALB/cJ which has a low intraocular pressure. Variant Thr-164 is found in strains C3H/HeJ and C57BL/6J, two strains which have a relatively high intraocular pressure.

NT3_HUMAN, P20783
POLYMORPHISM: Variant Glu-76 (frequently reported as Glu-63) was thought to be associated with severe forms of schizophrenia. This does not seem to be the case.

Variants & Keywords

P53_HUMAN, P04637

Keywords

3D-structure; Acetylation; Activator; Alternative splicing; Anti-oncogene; Apoptosis; Cell cycle; Disease mutation; DNA-binding; Glycoprotein; Li-Fraumeni syndrome; Metal-binding; Nuclear protein; Phosphorylation; Polymorphism; Transcription; Transcription regulation; Zinc.

Disease variants & Keywords

Keywords

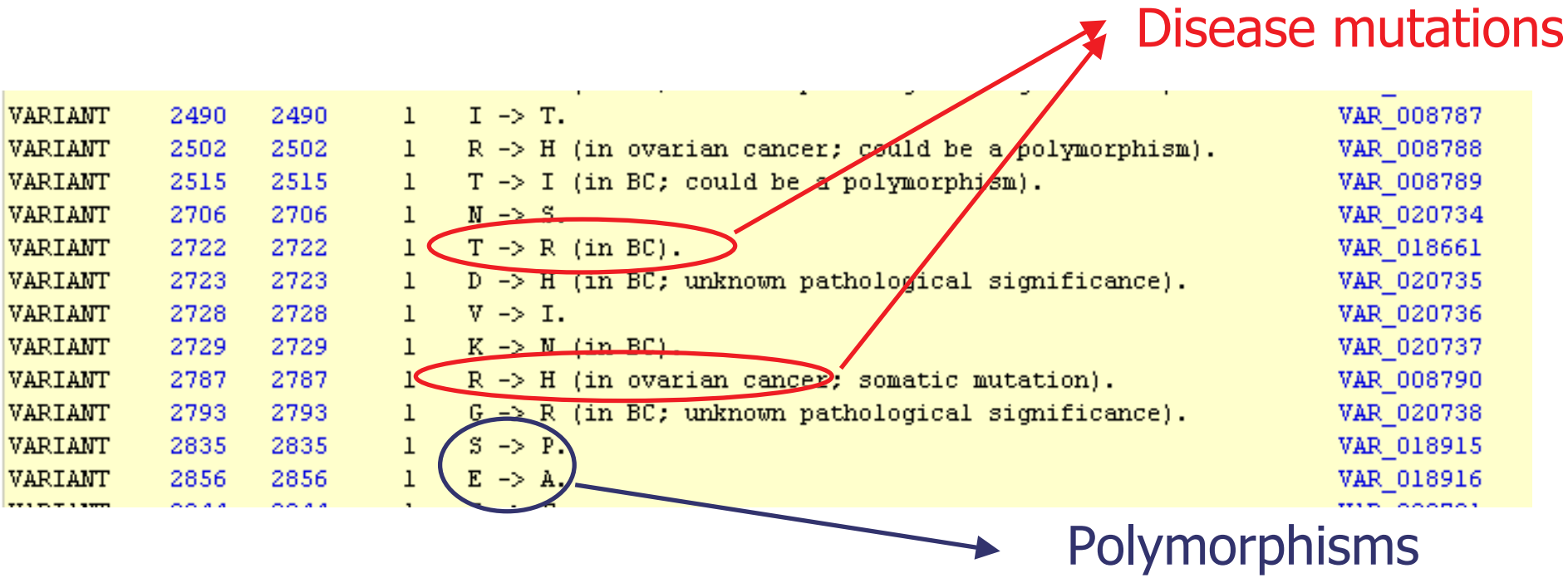
3D-structure; Acetylation; Activator; Alternative splicing; Anti-oncogene; Apoptosis; Cell cycle; Disease mutation; DNA-binding; Glycoprotein; Li-Fraumeni syndrome; Metal-binding; Nuclear protein; Phosphorylation; Polymorphism; Transcription; Transcription regulation; Zinc.

Disease-related keywords:

Alport syndrome, Alzheimer's disease, Autoimmune encephalomyelitis, Autoimmune uveitis, Bernard Soulier syndrome, Charcot-Marie-Tooth disease, Chronic granulomatous disease, Cockayne's syndrome, Cone-rod dystrophy, Deafness, Diabetes insipidus, Diabetes mellitus, Down's syndrome, Dwarfism, Epidermolysis bullosa, Gaucher disease, GM2-gangliosidosis, Hemophilia, Hereditary hemolytic anemia, Hereditary multiple exostoses, Hereditary nonpolyposis colorectal cancer, Hirschsprung disease, etc…

Variants & Features

BRCA2_HUMAN, P51587



P53_HUMAN, P04637

Variants & Features

VARIANT	94	94	1	S -> T (in a colon tumor).	VAR_005859 [3D]
VARIANT	110	110	1	R -> C (in a liver and an uterus tumor).	VAR_005860 [3D]
VARIANT	110	110	1	R -> L (in a liver tumor).	VAR_005861 [3D]
VARIANT	110	110	1	R -> P (in a breast tumor).	VAR_005862 [3D]
VARIANT	113	113	1	F -> C (in a lung tumor).	VAR_005863 [3D]
VARIANT	125	125	1	T -> M (in a lung tumor).	VAR_005864 [3D]
VARIANT	126	126	1	Y -> D (in a colorectal tumor).	VAR_005865 [3D]
VARIANT	126	126	1	Y -> N (in a leukemia and a lymphoma).	VAR_005866 [3D]
VARIANT	127	127	1	S -> F (in a lung tumor).	VAR_005867 [3D]
VARIANT	128	128	1	P -> S (in a breast tumor).	VAR_005868 [3D]
VARIANT	129	129	1	A -> D (in a sarcoma).	VAR_005869 [3D]
VARIANT	130	130	1	L -> R (in a liver tumor).	VAR_005870 [3D]
VARIANT	131	131	1	N -> K (in a colon tumor).	VAR_005872 [3D]
VARIANT	131	131	1	N -> S (in a liver tumor).	VAR_005871 [3D]
VARIANT	132	132	1	K -> M (in a sarcoma).	VAR_005873 [3D]
VARIANT	132	132	1	K -> Q (in a breast tumor).	VAR_005874 [3D]
VARIANT	133	133	1	M -> T (in LFS).	VAR_005875 [3D]
VARIANT	135	135	1	C -> F (in a colon tumor).	VAR_005877 [3D]
VARIANT	135	135	1	C -> S (in a colon tumor).	VAR_005876 [3D]
VARIANT	136	136	1	Q -> E (in a breast tumor).	VAR_005878 [3D]
VARIANT	136	136	1	Q -> K (in a colon tumor).	VAR_005879 [3D]
VARIANT	137	137	1	L -> Q (in a liver tumor).	VAR_005880 [3D]
VARIANT	138	138	1	A -> P (in a lung tumor).	VAR_005881 [3D]
VARIANT	139	139	1	K -> N (in a breast, an ovary tumor, a leukemia and a lymphoma).	VAR_005882 [3D]

Swiss-Prot variant: VAR_005868 in P04637

[General Information] [Information on the variant] [Structural Information on the variant] [References for the variant] [Cross references for the variant]

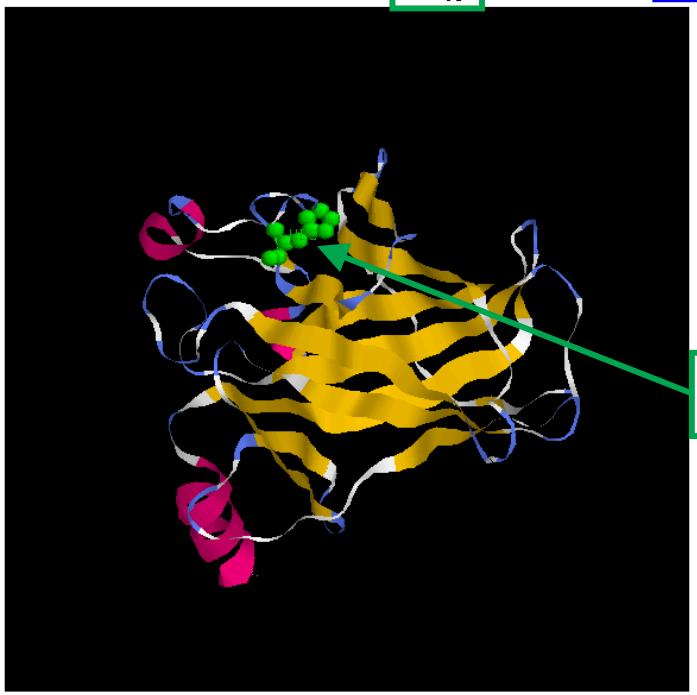
Note: Most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

General information	
Swiss-Prot ID (AC)	**P53_HUMAN (P04637)**
Gene symbol(s)	**Name: TP53** Synonym(s): P53
Chromosomal location	17p13.1
Protein name	Cellular tumor antigen p53
Length of the protein	393
Information on the variant	
FTId	**VAR_005868**
Amino acid position of the variant	128
Residue change	**From Pro (P) to Ser (S), P128S**
Status (Disease, polymorphism or unclassified)	Disease
Disease	A breast tumor
Comment	None
Structural information on the variant	
Location on the sequence	108 GFRLGFLHSGTAKSVTCTYS P ALNKMFCQLAKTCPVQLWVD 148 ↓ S
Protein features in neighborhood	**Key From To Length Description** CHAIN 1 393 393 Cellular tumor antigen p53 DNA_BIND 102 292 191 REGION 100 370 271 Interaction with HIPK1 (By similarity) REGION 116 292 177 Interaction with AXIN1 (By similarity) TURN 128 131 4
Residue conservation	Alignment from Blast search
Physico-chemical property	Change from medium size and hydrophobic (P) to small size and polar (S)
3D homology models	**Model Visualization** / **Template Structure**  ExPASy  AstexViewer  1TSRA [ExPASy / EBI-MSD]

PDB structure viewer

Other views : [Chime] [RasMol script] [PDB entry]

Wild type modelled structure of P04637



Model of wild type

Protein

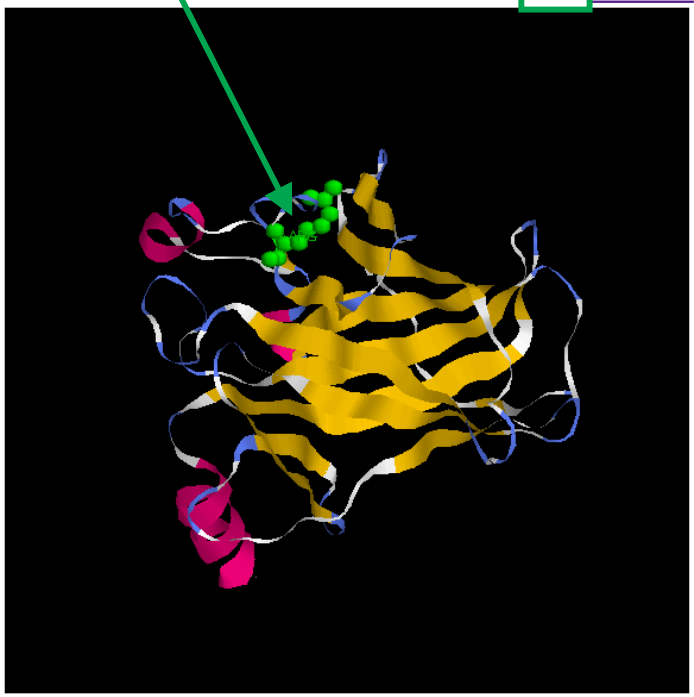
displayed as ribbons and colored in structure

Residue number 193

1 match displayed as ball-and-stick and colored in green

Model of variant

Modelled structure of variant VAR_005948: H193R



Protein

displayed as ribbons and colored in structure

Residue number 193

1 match displayed as ball-and-stick and colored in green

Swiss-Prot annotation and protein diversity

1. Alternative splicing.
2. Variants.
3. Post-translational modifications (PTM).

PTM definition

a post-translational modification or PTM is

a **modification of a polypeptidic chain** involving the **making or the breaking of covalent bond(s)** that occur **during** (co-translational class) **or after translation**.

PTMs influence or even define protein function

- phosphorylation and possibly GlcNAcylation and S-nitrosylation are a means of transducing extracellular signals to the inside of the cells.
- methylation has a role in nuclear protein import.
- lipid addition allows protein to membrane association (e.g. GPI-anchor, myristate, palmitate).
- intrachain disulfide bonds and N-glycosylation influence protein folding.
- interchain disulfide bonds bind subunits together.
- other PTMs are directly involved in the protein function, as for example the binding of cofactors (e.g. pyridoxal phosphate), or the synthesis of a cofactor by the modification of amino acids present in the protein (e.g. quinones).

3 categories of PTMs in Swiss-Prot

cleavage — initiator Met, signal and transit peptides, propeptides, complex processing, etc.

linkage

- simple chemical groups: phosphate, sulfate, methyl, hydroxyl, acetate, etc.
- complex molecules: N-, O- or C-linked glycans, lipids (e.g. palmitate, myristate, GPI)

cross-linking — disulfide bonds, thioester, thiolether bonds, etc.

3 locations for PTM data storage

- Comments (CC line type)
  CC -!- PTM: …
- Keywords (KW line type)
  KW …
- Feature table (FT line type)
  FT …

PTM keywords

cleavage keywords

Signal, Transit peptide, Protein splicing,
Cleavage on pair of basic residues

linkage keywords

Acetylation, ADP-ribosylation,
Amidation, D-amino acid, Formylation,
Glycoprotein, GPI-anchor, Hydroxylation,
Hypusine, Iodination, Myristate, Palmitate,
Phosphorylation, Prenylation, Sulfation

The keywords allow the retrieval of entries for proteins bearing the same modification.

PTM feature keys

cleavage FtKeys	INIT_MET, PROPEP, SIGNAL, TRANSIT
linkage FtKeys	MOD_RES, CARBOHYD, LIPID, BINDING
cross-link FtKeys	DISULFID, CROSSLNK

FtKey	position/range		FtDescription
INIT_MET	0	0	
BINDING	X	X	Pyridoxal phosphate.
DISULFID	X	Y	
MOD_RES	X	X	Glutamate methyl ester (Glu).
LIPID	X	X	GPI-anchor amidated glycine.
CARBOHYD	X	X	C-linked (Man).

Some statistics

Number of PTMs in Swiss-Prot release 47.3 (June 2005) for selected types of PTMs	all organisms			
	Pot./prob.	By sim.	Exp.	total
signal peptide	12246	2477	4688	19411
N-GlcNAc	55639	689	2162	58490
O-GalNAc	30	204	499	733
O-GlcNAc	8	97	52	157
phosphorylation	1344	6346	3424	11114
sulfation	225	200	149	574
myristate	109	310	122	531
GPI-anchor	372	112	43	527

50% of human entries describe at least one PTM
Total number of proteins < total number of PTMs
(but many of them are not yet annotated!)

Standardization of FT descriptions

http://www.expasy.org/sprot/userman.html#PTM_vocabularies

- MOD_RES, CROSSLNK, LIPID have been standardized, to use controlled vocabularies
- FT description shows the modified amino acid, not the name of the process:

PHOSPHORYLATION

OLD

Phosphocysteine
4-aspartylphosphate
Phosphohistidine
Tele-phosphohistidine
Pros-phosphohistidine
Phosphoserine
Phosphothreonine
Phosphotyrosine

NEW

Keyword « Phosphorylation » allows to retrieve all relevant entries

PTM comments : some examples

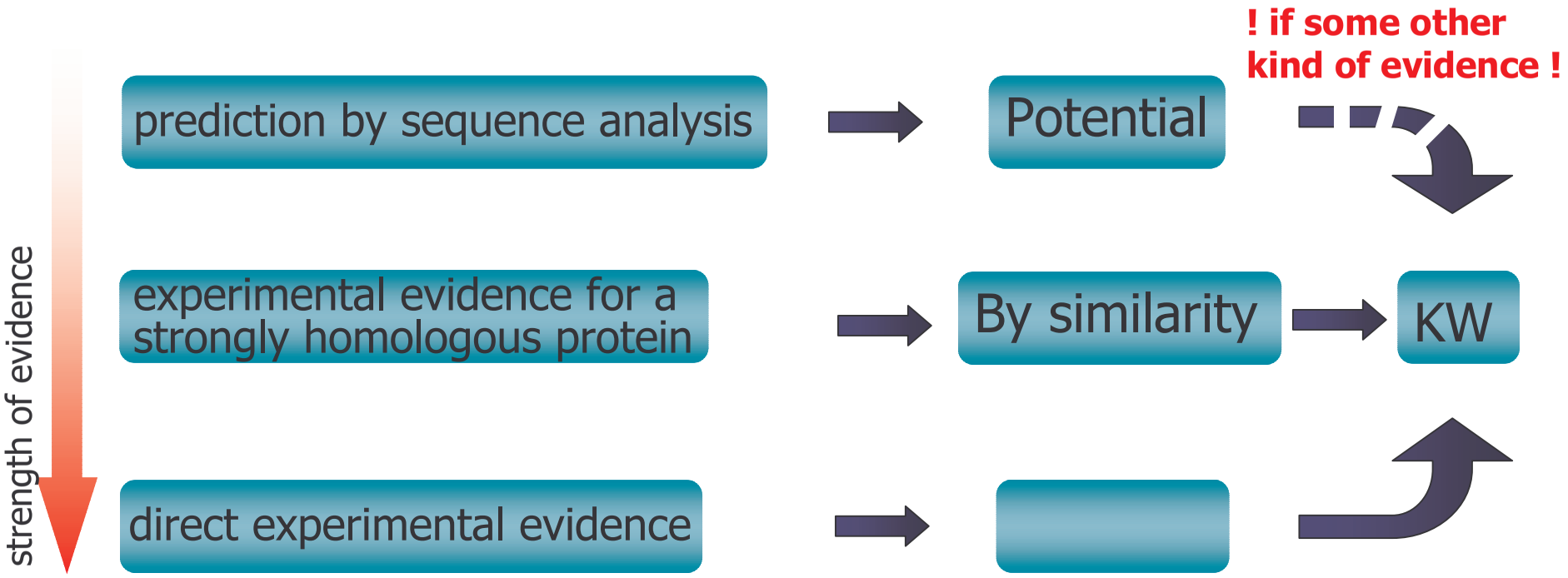
The comment (CC line-type) topic '**PTM**' is used to present detailed information on PTMs that cannot be represented in the feature table.

- The N-terminus is blocked.
- Phosphorylated by various PKC isozymes.
- Phosphorylated at multiple sites by different protein kinases and each phosphorylation event selectively modulates the protein's functions.
- Phosphorylation on Tyr-660 reduces the ability of 4.1 to promote the assembly of the spectrin/actin/4.1 ternary complex.
- O-glycosylated; contains N-acetylglucosamine side chains in the C-terminal domain.
- Sulfated.

Our sources of data

- PTM prediction tools (potential evidence)
- research articles, reviews (experimental, potential evidence)
- personal communications (experimental evidence)
- direct submissions of information to Swiss-Prot (experimental evidence)

≠ data sources, ≠ qualifiers



Swiss-Prot considers both data obtained experimentally and predicted information and makes a clear distinction in its annotation.

ExPASy Proteomics tools

http://www.expasy.org/tools/

The tools marked by are local to the ExPASy server. The remaining tools are developed and hosted on other servers.

[Protein identification and characterization] [DNA -> Protein] [Similarity searches] [Pattern and profile searches] [Post-translational modifi
[Primary structure analysis] [Secondary structure prediction] [Tertiary structure] [Sequence alignment] [Phylogenetic analysis

Protein identification and characterization

Identification and characterization with peptide mass fingerprinting data

- Aldente - Identify proteins with peptide mass fingerprinting data. A new, fast and powerful tool that takes advantage of h recalibration and outlier exclusion
- FindMod - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and characterize the protein of interest.
- FindPept - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into ac post-translational modifications (PTM) and protease autolytic cleavage
- GlycoMod - Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses oligosaccharides and for glycopeptides)
- Mascot - Peptide mass fingerprint from Matrix Science Ltd., London
- PepMAPPER - Peptide mass fingerprinting tool from UMIST, UK
- PFMUTS - Shows the possible single and double mutations of a peptide fragment from MALDI peptide mass fingerprinting
- ProFound - Search known protein sequences with peptide mass information from Rockefeller and NY Universities [or from Gen
- ProteinProspector - UCSF tools for peptide masses data (MS-Fit, MS-Pattern, MS-Digest, etc.)

Identification and characterization with MS/MS data

- Popitam new - Identification and characterization tool for peptides with unexpected modifications (e.g. post-translational spectrometry
- Phenyx - Protein and peptide identification/characterization from MS/MS data from GeneBio, Switzerland
- Mascot - Sequence query and MS/MS ion search from Matrix Science Ltd., London
- OMSSA - MS/MS peptide spectra identification by searching libraries of known protein sequences
- PepFrag - Search known protein sequences with peptide fragment mass information from Rockefeller and NY Universities [or fi

- Use annotation in Swiss-Prot and TrEMBL (preprocessing, PTMs, etc.)
- Hyper-links between tools and databases

Identification:
Aldente,
TagIdent,
AAcompIdent,
MultiIdent

Characterization:
FindMod,
GlycoMod,
FindPept

Analysis:
PeptideMass,
GlycanMass,
BioGraph,
PeptideCutter
ProtScale,
ProtParam


ONCE YOU UNFOLD ONE OF THESE THINGS, ITS NEVER THE SAME.
Ethical Questions
Medical Dilemmas
MAP OF THE HUMAN GENOME
Legal Tangles
Privacy Concerns
Moral Issues
Insurance Applications

Protein families and domains, PROSITE and InterPro

Sequence patterns and profiles

- Detection of protein sequence families or domain using sequence patterns;
- Detection of DNA sequence elements or protein families/domains using weight matrices (or profiles), and Hidden Markov Models.

Identification of protein domains and families

- There are two non-exclusive approaches for the determination of the function of an uncharacterized protein:
  - Comparison with a complete sequence database
  - Scanning a database of patterns and profiles
- Most proteins can be grouped into families. Proteins belonging to a particular family share functional attributes and are derived from a common ancestor;
- Some regions in the sequence are more conserved than others during evolution because they are important for the function or the structure of the protein;
- Like fingerprints for police identification, signatures built out of sequence patterns or profiles can be used to formulate hypotheses about the function of uncharacterized proteins.

How many protein families? [1/2]

Not an easy question to answer:

- The number of different proteins expressed by a genome range from about 500 (Mycoplasma) to probably 30'000 (Human) and a lot more in plants. But these numbers do not translate to an equivalent number of families. The increase in protein numbers is mainly due to multigenic families;
- According to Chothia (1992): about 1'000 protein families based on structural criteria (folds);
- According to Claverie, States and Lipman (1993): about 600 ACS (**A**ncestral **C**onserved **S**equences), families found in all phyla. The other families are specific for one or more phyla;
- Number of eukaryotic protein modules according to Bork (1992): about 200. These are modules that make up "Lego" proteins. Examples: kringle, IG-like, EGF, FN-III, etc.;

How many protein families? [2/2]

- Each phylum and in some cases each species seems to contain specific protein families. Examples: fungal-specific proteins; plant-specific, C.elegans-specific; etc.;
- The number of protein families is dependent on how far sequence analysis can go back in evolution. At the extreme there should only be one protein family!
- Structural criteria generally allow to go the farthest back in time, but it is not proven that similar folds equate with a common ancestor;
- So any estimation of the number of protein families or domains is dependent on the criteria used to define what is a family or a domain;
- We estimate that an "operational" protein classification system requires probably some 3’000 to 5’000 "definitions".

http://www.ebi.ac.uk/integr8/ProteomeAnalysisAction.do?orgProteomeID=25

InterPro top 30 entries for *H. sapiens*

InterPro	Proteins matched (Proteome coverage)	Rank *	Name
IPR007110	1176(3.1%)	1	Immunoglobulin-like
IPR007087	1055(2.8%)	2	Zinc finger, C2H2-type
IPR003599	977(2.6%)	3	Immunoglobulin subtype
IPR000276	895(2.4%)	4	Rhodopsin-like GPCR superfamily
IPR011009	883(2.3%)	5	Protein kinase-like
IPR000719	808(2.1%)	6	Protein kinase
IPR011993	596(1.6%)	7	Pleckstrin homology-type
IPR013106	591(1.6%)	8	Immunoglobulin V-set
IPR000725	508(1.3%)	9	Olfactory receptor
IPR007086	489(1.3%)	10	Zinc finger, C2H2-subtype
IPR001909	463(1.2%)	11	KRAB box
IPR011992	436(1.1%)	12	EF-Hand type
IPR002110	424(1.1%)	13	Ankyrin
IPR008271	424(1.1%)	13	Serine/threonine protein kinase, active site
IPR001849	410(1.1%)	15	Pleckstrin-like
IPR012677	409(1.1%)	16	Nucleotide-binding, alpha-beta plait
IPR001680	404(1.1%)	17	WD-40 repeat

How to identify protein families or domains

PATTERNS = regular expressions;

PROFILES or **WEIGHT MATRICES** = two-dimensional tables of position specific match-, gap-, and insertion-scores, derived from aligned sequence families;

Hidden Markov Models (HMM) = probabilistic models.

How to judge the quality of a discriminator

Two parameters can be computed to qualify the quality of a sequence discriminator: **precision** and **recall**

False positives = known false hits

False negatives = known missed hits

Precision = true hits / (true hits + false positives)

Recall = true hits / (true hits + false negatives)

Sensitivity: does it pick up *all* members of the family?

Specificity: does it find *only* true members of the family?

A regular expression (RE) expresses how a computer program should look for a specified pattern in text

'^' A circumflex at the beginning of the pattern matches the line beginning.

'$' A dollar-sign at the end of the pattern matches the end of a line.

'.' A period matches any character except for 'new-line'.

'*' An expression followed by an asterisk matches zero or more occurrences of that expression : 'fo*' matches 'f', 'fo', 'foo', etc.

'+' An expression followed by a plus sign matches one or more occurrences of that expression : 'fo+' matches 'fo', 'foo', etc.

'[]' A string enclosed in square brackets matches any character in that string, but no others. If the 1st character is a circumflex it will match any character **except** the characters in the string. Example: '[xyz]' matches matches 'xx' and 'zyx'; '[^xyz]' matches 'abc' but not 'axb'.

A range of characters may be specified by two characters separated by '-'. Example: [a-z] matches any word containing any lower case character.

Regular expressions - examples

C[AOU]T- Would match : **CAT, COT, CUT** only

C.T - Would match : **CAT, CaT, CIT, C:T** NOT CT

CA?T - Would match : **CT, CAT** only

C+T - Would match **CT, CCT, CCCCT**

C(HE)?A[TP] - Would match : **CHEAT, CAT, CHEAP, CAP** only

Sequence patterns

- From a multiple sequence alignment one can derive a sequence pattern;
- Useful to detect protein belonging to a specific family or a protein domain; much less useful at the DNA level due to the small alphabet (4 letters) and the low sequence conservation of DNA sequence elements;
- Examples of sequence patterns:

  Serine active site of trypsin-type proteases: Gly-Asp-**Ser**-Gly-Gly

  Zinc finger C2H2-type: Cys-x(2,4)-Cys-x(12)-His-x(3,5)-His

  Chitin-binding domain signature: C-x(4)-C-C-S-x(2)-G-x-C-G-x(4)-[FYW]-C

PROSITE pattern conventions

1. The standard IUPAC one-letter codes for the amino acids are used.
2. The symbol `x' is used for a position where any amino acid is accepted.
3. Ambiguities are indicated by listing the acceptable amino acids for a given position, between square parentheses `[ ]'. For example: [ALT] stands for Ala or Leu or Thr.
4. Ambiguities are also indicated by listing between a pair of curly brackets `{ }' the amino acids that are not accepted at a given position. For example:{AM} stands for any amino acid except Ala and Met.
5. Each element in a pattern is separated from its neighbor by a '-'.
6. Repetition of an element of the pattern can be indicated by following that element with a numerical value or, if it is a gap ('x'), by a numerical range between parentheses. Examples:

   x(3) corresponds to x-x-x

   x(2,4) corresponds to x-x or x-x-x or x-x-x-x

   A(3) corresponds to A-A-A

   Note: You can only use a range with 'x', i.e. A(2,4) is not a valid pattern element.
7. When a pattern is restricted to either the N- or C-terminal of a sequence, that pattern either starts with a `<' symbol or respectively ends with a `>' symbol.

How to build protein sequence patterns 1/2

- One starts with a multiple sequence alignment;
- Choose a region with well-conserved residues
- If something is known about the sequence, it is useful to consider residues and regions thought/proved to be important to the biological function of that group of proteins. These biologically significant regions or residues are generally:
  - Enzyme catalytic sites;
  - Prosthetic group attachment sites (heme, PLP, biotin, etc.);
  - Amino acids involved in binding a metal ion;
  - Cysteines involved in disulfide bonds;
  - Regions involved in binding a molecule (ATP, calcium, DNA etc.) or a protein.

How to build protein sequence patterns 2/2

- A first pattern is built from the most conserved residues. It is used to scan the database;
- If it picks up too many false positives, it is modified to make it more stringent;
- The difficulty resides in achieving a pattern which does not pick up too many false positives yet does not miss too many sequences (false negatives). In practice it means optimizing both the recall and the precision;
- In some cases this result can not be achieved and an optimal sequence pattern cannot be built.

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

A database of protein families and domains using two kinds of motif descriptors:

Patterns or regular expressions :

- User friendly (easy to understand and to use)
- Well designed for the detection of biologically meaningful sites such as residues playing a structural or functional role
- Can be used to scan a protein database in reasonable time on any computer

Generalized profiles or weight matrices :

- Well adapted to cover the full length of the protein or domain
- Are able to detect highly divergent families or domains with only few well conserved positions

PROSITE - example pattern entry

```
ID   PPASE; PATTERN.
AC   PS00387;
DT   NOV-1990 (CREATED); NOV-1997 (DATA UPDATE); MAY-2006 (INFO UPDATE).
DE   Inorganic pyrophosphatase signature.
PA   D-[SGDN]-D-[PE]-[LIVMF]-D-[LIVMGAC].
NR   /RELEASE=50.1,223100;
NR   /TOTAL=162(162); /POSITIVE=109(109); /UNKNOWN=1(1); /FALSE_POS=52(52);
NR   /FALSE_NEG=8; /PARTIAL=2;
CC   /TAXO-RANGE=A?EP?; /MAX-REPEAT=1;
CC   /SITE=1,magnesium; /SITE=3,magnesium; /SITE=6,magnesium;
CC   /VERSION=1;
DR   Q889M7, IPYR1_PSESM, T; Q9H2U2, IPYR2_HUMAN, T; Q91VM9, IPYR2_MOUSE, T;
DR   Q87WD6, IPYR2_PSESM, T; P87118, IPYR2_SCHPO, T; P28239, IPYR2_YEAST, T;
DR   Q9YBA5, IPYR_AERPE , T; Q8UC37, IPYR_AGRT5 , T; P80562, IPYR_ANASP , T;
DR   O05545, IPYR_GLUOX , N; Q72MG4, IPYR_LEPIC , N; Q8EZ21, IPYR_LEPIN , N;
DR   Q8PH18, IPYR_XANAC , N; Q8P5M4, IPYR_XANCP , N;
DR   Q06305, AER3_AERHY , F; Q06303, AER4_AERHY , F; Q06306, AER5_AERHY , F;
DR   P09167, AERA_AERHY , F; Q9X4Y1, AGPA_RHIME , F; Q44257, CBAB_COMTE , F;
3D   1FAJ; 1HUJ; 1HUK; 1IGP; 1INO; 1IPW; 1JFD; 1M38; 1MJW; 1MJZ; 1OBW; 1QEZ;
3D   1WGI; 1WGJ; 1YPP; 2EIP; 2PRD; 8PRK;
DO   PDOC00325;
//
```

Performance diagnostic

List of matches

PROSITE html view

http://www.expasy.org/prosite/PS00387

NiceSite View of: PS00387

General information about the entry	
Entry name	**PPASE**
Accession number	**PS00387**
Entry type	PATTERN
Date	NOV-1990 (CREATED); NOV-1997 (DATA UPDATE); MAY-2006 (INFO UPDATE).
PROSITE documentation	PDOC00325
Name and characterization of the entry	
Description	Inorganic pyrophosphatase signature.
Pattern	D-[SGDN]-D-[PE]-[LIVMF]-D-[LIVMGAC].

Numerical results

- UniProtKB/Swiss-Prot release number: **50.1**, total number of sequence entries in that release: **223100**.
- Total number of hits in UniProtKB/Swiss-Prot: **162 hits in 162 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **109 hits in 109 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **1 hits in 1 different sequences**
- Number of false hits (on unrelated proteins): **52 hits in 52 different sequences**
- Number of known missed hits: **8**
- 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: **2**
- Precision (true hits / (true hits + false positives)): **67.70 %**
- Recall (true hits / (true hits + false negatives)): **93.16 %**

Comments

- Taxonomic range: **Archaebacteria, Eukaryotes, Prokaryotes (Bacteria)**
- Maximum known number of repetitions of the pattern in a single protein: **1**
- `Interesting' site in the pattern: **1,magnesium**
- `Interesting' site in the pattern: **3,magnesium**
- `Interesting' site in the pattern: **6,magnesium**
- VERSION: **1**

Cross-references

True positive hits:

IPYR1_PSESM (Q889M7), IPYR2_HUMAN (Q9H2U2), IPYR2_MOUSE (Q91VM9),
IPYR2_PSESM (Q87WD6), IPYR2_SCHPO (P87118), IPYR2_YEAST (P28239),
IPYR_AERPE (Q9YBA5), IPYR_AGRT5 (Q8UC37), IPYR_ANASP (P80562),
IPYR_AQUAE (O67501), IPYR_AQUPY (Q8GQS5), IPYR_ARATH (P21216),
IPYR_ASHGO (Q757J8), IPYR_BACHD (Q9KCG7), IPYR_BACP3 (P19514),
IPYR_BACST (O05724), IPYR_BARBA (P51064), IPYR_BOVIN (P37980),

http://www.expasy.org/tools/scanprosite/

ScanProsite

The ScanProsite tool [Help] allows to scan protein sequence(s) (either from UniProt Knowledgebase (Swiss-Prot/TrEMBL) or PDB or provided by the user) for the occurrence of patterns, profiles and rules (motifs) stored in the PROSITE database, or to search protein database(s) for hits by specific motif(s) [Reference / Download ps_scan, the standalone version]. The program PRATT can be used to generate your own patterns. You may either:

- Enter one or more PROSITE accession numbers and/or patterns [1 by line] to search the UniProt Knowledgebase (Swiss-Prot/TrEMBL) and/or PDB databases, **OR**
- Enter one or more sequences [raw, Swiss-Prot or fasta format] and/or UniProt Knowledgebase (Swiss-Prot/TrEMBL) accession numbers and/or PDB accession numbers [1 by line] to be scanned with all patterns, profiles, rules in PROSITE, **OR**
- Fill in both fields to find all occurrences of specified motifs in specified sequences.

Protein(s) to be scanned:

Enter one or more Swiss-Prot/TrEMBL accession number(s) [AC] (e.g. **P00747**) and/or sequence identifier(s) [ID] (e.g. **ENTK_HUMAN**) , and/or PDB identifier, and/or paste **your own protein sequence(s)** in the box below:
(leave this box blank to scan PROSITE entrie(s) against selected protein databases)

Clear

General options:

- [x] Exclude motifs with a high probability of occurrence
- [ ] Show low level score
- [ ] Do not scan profiles [User Manual]

Show only sequences with at least [ ] hit(s)
Maximum of matched sequences [1000]

PROSITE pattern(s)/profile(s) to scan for:

Enter one or more PROSITE accession number(s) (e.g. **PS50240**), and/or identifier(s) (e.g. **CHEB**), and/or type **your pattern(s)** in PROSITE format in the box below:
(leave this box blank to scan sequence(s) against the entire PROSITE database)

and specify your search limits (only used if no protein data specified) :

- **Protein database(s):** [x] Swiss-Prot [ ] TrEMBL [ ] PDB databases
  [x] including splice variants
  randomize databases [no] (only with patterns, see help)
- Taxonomic lineage (OC) / species (OS) filter: [ ]
  (see NEWT Taxonomy ; separate multiple taxa/species with a semicolon, e.g. *Eukaryota; Escherichia coli;* . Does not work on PDB sequences.)
- Description (DE) filter: [ ] e.g. *protease*

pattern options:

Allow at most [1] X sequence characters to match a conserved position in the pattern
match mode [greedy, overlaps, no includes] (for patterns, see help)

ScanProsite Results: Pattern against UniProtKB/Swiss-Prot

hits by patterns: [149 hits (by 1 pattern) on 149 sequences]

Hits by **USERPAT1** :
Pattern: **D - [SGDN] - D - [PE] - [LIVMF] - D - [LIVMGAC]**
Approximate number of expected random matches [Ref: PMID 11535175] in Swiss-Prot release 41 (122564 sequences): 19

Q06305
(AER3_AERHY) (492 aa) individual view

Aerolysin 3 precursor (Hemolysin 3). *Aeromonas hydrophila*

118 - 124: DGDEVDV

Q06303
(AER4_AERHY) (492 aa) individual view

Aerolysin 4 precursor (Hemolysin 4). *Aeromonas hydrophila*

118 - 124: DGDEVDV

Q06306
(AER5_AERHY) (485 aa) individual view

Aerolysin 5 precursor (Hemolysin 5). *Aeromonas hydrophila*

118 - 124: DGDEVDV

P09167
(AERA_AERHY) (493 aa) individual view

Aerolysin precursor. *Aeromonas hydrophila*
Hits on PDB 3D structures: [1PRE-A,1PRE-B]

.....and many more matches....

PROSITE - profiles

Profile or weight matrix: table of position-specific amino acid weights and gap costs, used to calculate a *similarity score* for any alignment between a profile and a sequence. An alignment with a similarity score higher than or equal to a given *cut-off value* constitutes a motif occurrence.

More robust and sensitive than patterns due to **discriminatory weights** not only for the residues already found at a given position of a motif, but also for those not yet found. The weights for those not yet found are extrapolated from the observed amino acid compositions using empiric knowledge about amino acid substitutability.

Weight matrices (profiles) (1/5)

First step: start with a multiple sequence alignment (protein or DNA)

1	2	3	4	5	6	7
A	S	T	A	M	P	V
A	T	S	L	M	V	T
S	S	S	L	M	L	T
A	T	P	A	M	S	S
A	T	A	L	L	S	A

Weight matrices (profiles) (2/5)

2nd step: count the number of occurrences of the different amino acids (or bases) at each position of the alignment

1	2	3	4	5	6	7
4A	3T	2S	3L	4M	2S	1V
1S	2S	1T	2A	1L	1L	2T
		1A			1V	1S
		1P			1P	1A

Weight matrices (profiles) (3/5)

3rd step: derivation of the preliminary matrix

	1	2	3	4	5	6	7
A	0.8	0	0.2	0.4	0	0	0.2
L	0	0	0	0.6	0.2	0.2	0
M	0	0	0	0	0.8	0	0
V	0	0	0	0	0	0.2	0.2
P	0	0	0.2	0	0	0.2	0
S	0.2	0.4	0.4	0	0	0.4	0.2
T	0	0.6	0.2	0	0	0	0.4

The matrix size is N x M where N(columns) is the number of positions in the alignment and M(rows) is either 20 (for a protein matrix) or 4 (for DNA/RNA)

Weight matrices (profiles) (4/5)

Subsequent steps:

- Normalization (taking into account the average amino-acid composition);
- Replacing zero values by a small, but non-zero value;
- Using logarithms so as to speed up computation time;
- Definition of a cut-off value.

Once all this is done, the weight matrix can be used to scan the sequences in a database

Weight matrices (profiles) (5/5)

Important refinements

- Allowing gaps in the matrix; this is not trivial, but necessary to a real-life application;
- Using a weighted multiple alignment. If this is not done then the matrix will be biased toward the most represented subgroups of sequences.

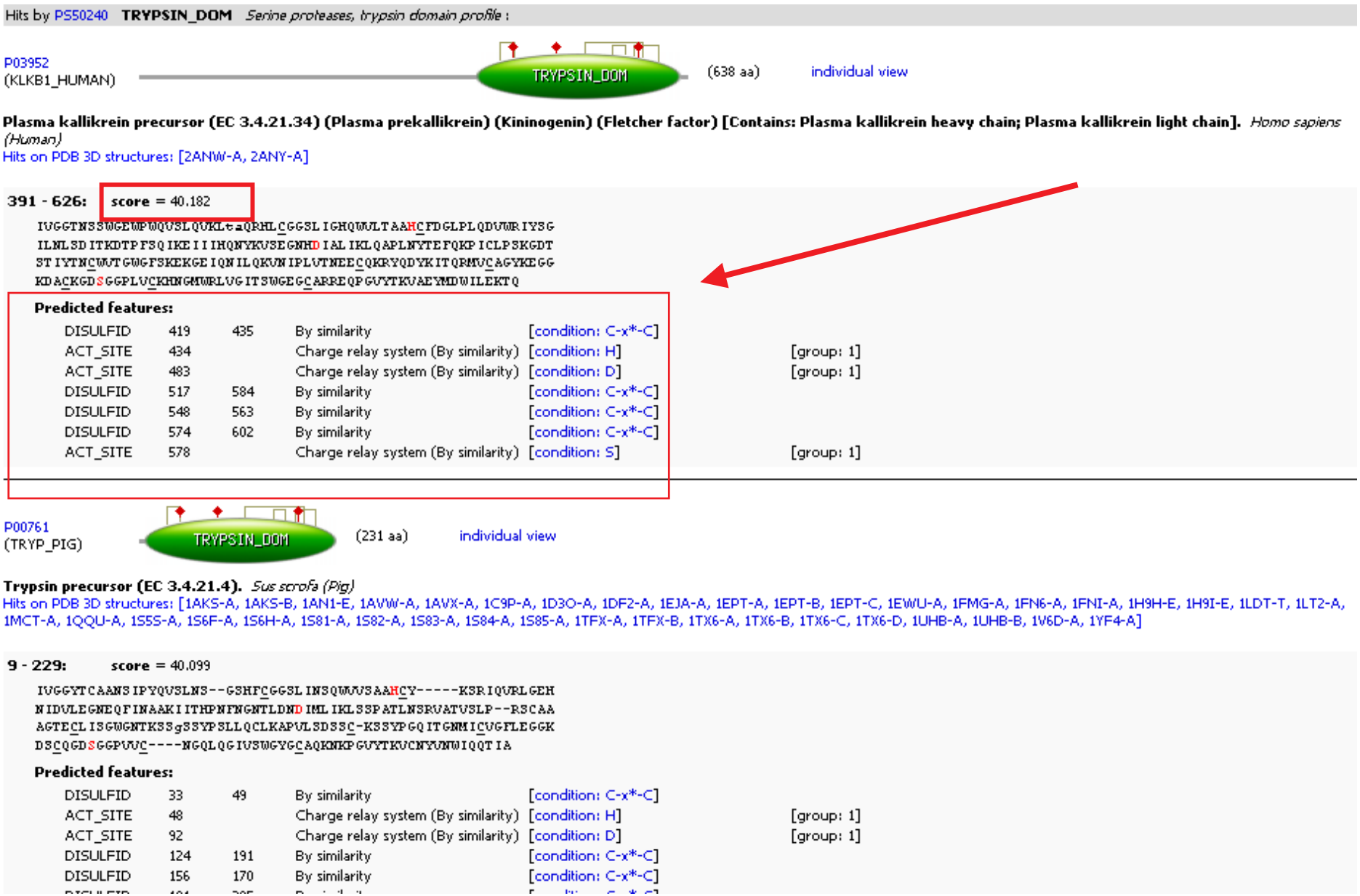
An example PROSITE profile entry: http://www.expasy.org/prosite/PS50240

NiceSite View of: PS50240

General information about the entry	
Entry name	**TRYPSIN_DOM**
Accession number	**PS50240**
Entry type	MATRIX
Date	DEC-2001 (CREATED); DEC-2001 (DATA UPDATE); MAY-2006 (INFO UPDATE).
PROSITE documentation	PDOC00124
Name and characterization of the entry	
Description	Serine proteases, trypsin domain profile.

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=234;
/DISJOINT: DEFINITION=PROTECT; N1=6; N2=229;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=0.0169; R2=0.00836256; TEXT='-LogE';
/CUT_OFF: LEVEL=0; SCORE=1134; N_SCORE=9.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=775; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: M0=-9; D=-20; I=-20; B1=-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:         B1=0; BI=-105; BD=-105;
/M: SY='I'; M= -7,-29,-26,-37,-28,  0,-35,-29, 41,-28, 20, 17,-22,-22,-21,-27,-19, -9, 29,-21, -2,-28;
/M: SY='V'; M= -3,-27,-15,-30,-25,  0,-31,-26, 26,-20, 10,  9,-24,-25,-23,-20,-13, -3, 32,-22, -4,-26;
/M: SY='G'; M=  0, -2,-27, -5,-15,-28, 54,-15,-35,-16,-29,-20,  8,-19,-16,-17,  3,-15,-28,-24,-28,-15;
/M: SY='G'; M=  0, -7,-29, -7,-16,-29, 60,-18,-37,-18,-28,-18,  1,-19,-17,-18,  0,-18,-29,-21,-28,-16;
/I:         I=-3; MD=-17;
/M: SY='Q'; M= -4, -2,-19, -4,  4,-13,-16,  0,-15, -1,-14, -7,  0,-14,  5,  0,  4,  2,-12,-23, -5,  4; D=-3;
/I:         I=-3; MI=0; MD=-17; IM=0; DM=-17;
/M: SY='E'; M= -5,  4,-24,  7, 13,-22,-14, -7,-18,  0,-17,-13,  0, -3,  1, -5,  0, -4,-15,-29,-16,  6; D=-3;
/I:         I=-3; DM=-17;
/M: SY='A'; M= 18, -9, 10,-17,-13,-18,-13,-20,-10,-14,-12,-10, -8,-15,-11,-19,  6,  5, -1,-29,-18,-13; D=-3;
/I:         I=-3; DM=-17;
/M: SY='B'; M= -3,  2,-22,  1,  2,-19,-15, -5,-15, -1,-14,-10,  2,-10,  0, -3,  2,  2,-12,-27,-11,  0;
/M: SY='P'; M= -3,-15,-28,-15, -3,-16,-21,-14, -3, -6, -9, -4,-12,  8, -4, -7, -8, -8, -7,-21,-12, -6;
```

ScanProsite results: PROSITE profile against Swiss-Prot



Functionally and structurally relevant residues in PROSITE motif descriptors

A new concept to extract more information from profiles

Principle :

- Combining the advantages of profiles (high sensitivity) and patterns (position-specific information)
- Tagging of amino acids at precise positions in the profile and checking their presence in the matched sequence

Summary of advantages of patterns and profiles

Patterns or regular expressions :

- User-friendly (easy to understand and to use)
- Well-designed for the detection of biologically meaningful sites such as residues playing a structural or functional role
- Can be used to scan a protein database in reasonable time on any computer

Generalized profiles or weight matrices :

- Well adapted to cover the full length of the protein or domain
- Are able to detect highly divergent families or domains with only few well-conserved positions

ProRule: A help for function prediction and annotation

Sigrist et al.: Bioinformatics 21:4060-4066(2005)

Aim :

- Provide users with biologically meaningful functional and structural information:
  - active sites,
  - post-translational modification sites,
  - binding sites,
  - disulfide bonds,
  - transmembrane regions.

- Help the Swiss-Prot annotation and provide enhanced homogeneity:
  - domain name and boundaries,
  - keywords and linked GO terms,
  - EC numbers,
  - false negative PROSITE patterns.

Annotation rule PRU00274

ProRule is a database containing additional information about PROSITE profiles. It is used to help function prediction and annotation of the Swiss-Prot knowledgebase. The ProRule database uses the UniRule format that is in the form of Swiss-Prot lines and is identical for all types of rules created to annotate Swiss-Prot, like the HAMAP family rules. Each rule is triggered by a PROSITE profile and contains information linked to the domain or protein family covered by the profile. This information can be general, e.g. always associated with the domain or protein family, or conditional, depending on the presence of particular residues in functionally or structurally critical positions.

[?] General information about the entry

Accession	PRU00274
Dates	12-DEC-2003 (Created) 12-MAY-2005 (Last updated, Version 13)
Data class	Domain
Predictors	PROSITE; PS50240; TRYPSIN_DOM
Names	Serine proteases, trypsin domain
Function	Cleaves preferentially: Arg-|-Xaa, Lys-|-Xaa
Description	+ (EC 3.4.21.-) **[case <FTGroup:1>]**

[?] Comments

- ***SIMILARITY:*** Belongs to the peptidase S1 family.
- ***SIMILARITY:*** Contains # peptidase S1 domain.

[?] Cross-references

PROSITE	PS00134; TRYPSIN_HIS; 1 **[case <FTGroup:1>]** PS00135; TRYPSIN_SER; 1 **[case <FTGroup:1>]** PS00134; TRYPSIN_HIS; 0-1 **[case not(<FTGroup:1>)]** PS00135; TRYPSIN_SER; 0-1 **[case not(<FTGroup:1>)]**

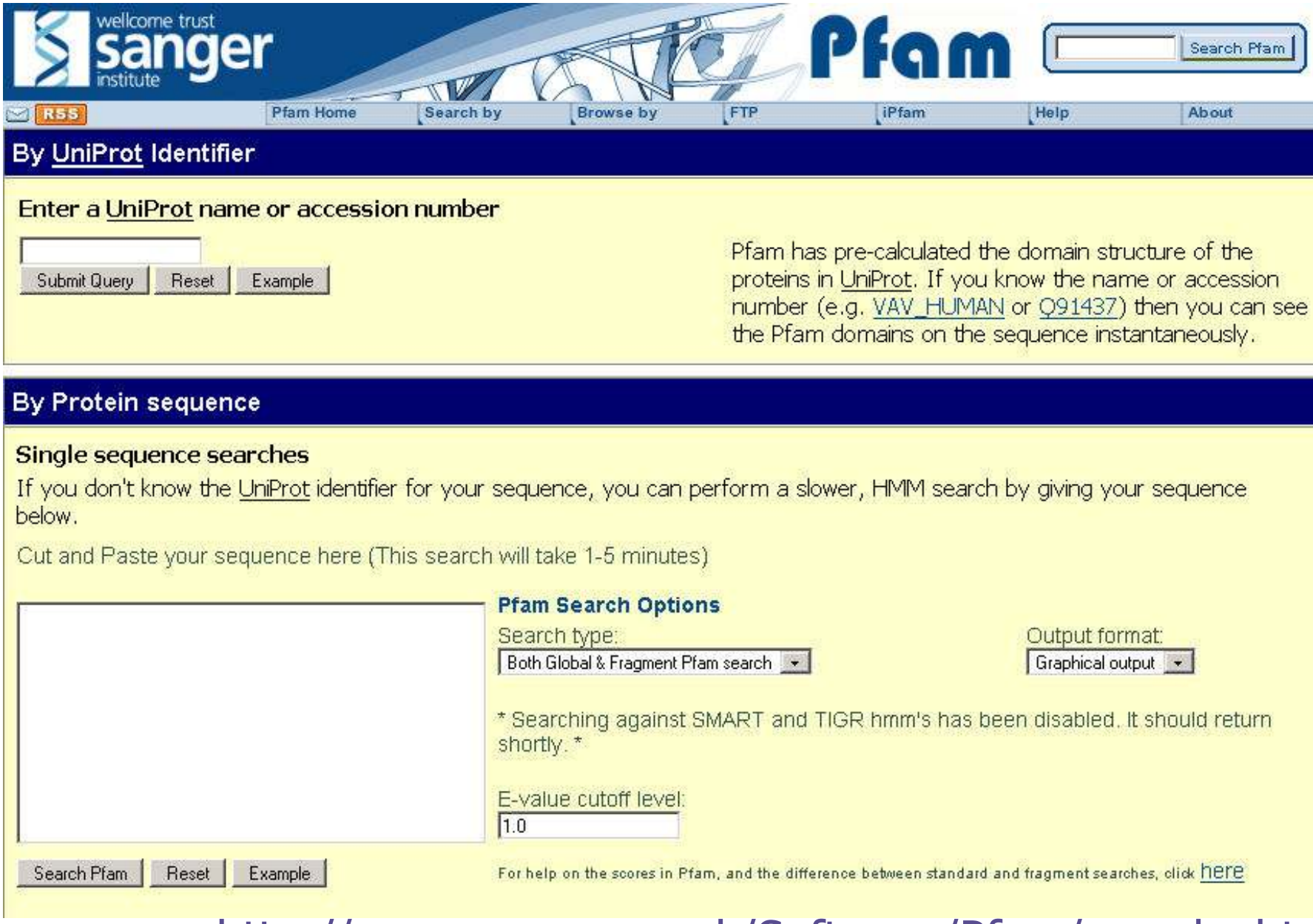
[?] Features

Key	From	To	Description	Condition	[?] case
From: PS50240					
DOMAIN	from	to	Peptidase S1 #		
ACT_SITE	42	42	Charge relay system *(By similarity)*	H	FTGroup[1]
ACT_SITE	91	91	Charge relay system *(By similarity)*	D	FTGroup[1]
ACT_SITE	186	186	Charge relay system *(By similarity)*	S	FTGroup[1]
DISULFID	27	43	*By similarity*	C-x*-C	
DISULFID	125	192	*By similarity*	C-x*-C	
DISULFID	156	171	*By similarity*	C-x*-C	
DISULFID	182	210	*By similarity*	C-x*-C	

Hidden Markov Models (HMM)

- A powerful probabilistic method;
- The advantage of HMM is that you can use them to both build the multiple sequence alignment and to detect if a protein belongs to a family or a domain;
- For a review see:

  Bioinformatics 14:755-763(1998)



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

PROSITE - tools

http://www.expasy.org/tools/#pattern :

- ScanProsite
- myHits MotifScan
- … and many others to scan a sequence against PROSITE or a pattern/profile against a protein sequence database

- Interpro Scan:

http://www.ebi.ac.uk/interpro/scan.html

Other methods

Pfam, TIGRFAMs, SMART: Hidden Markov Models (HMM), Probabilistic model;

PRINTS: "Unweighted" matrix; protein fingerprints

BLOCKS: Weight matrix derived from ungapped alignments;

PIR SuperFamily: classification system based on evolutionary relationship of whole proteins

ProDom: automatic compilation of homologous domains based on recursive PSI-BLAST searches.

The INTERPRO project
www.ebi.ac.uk/interpro

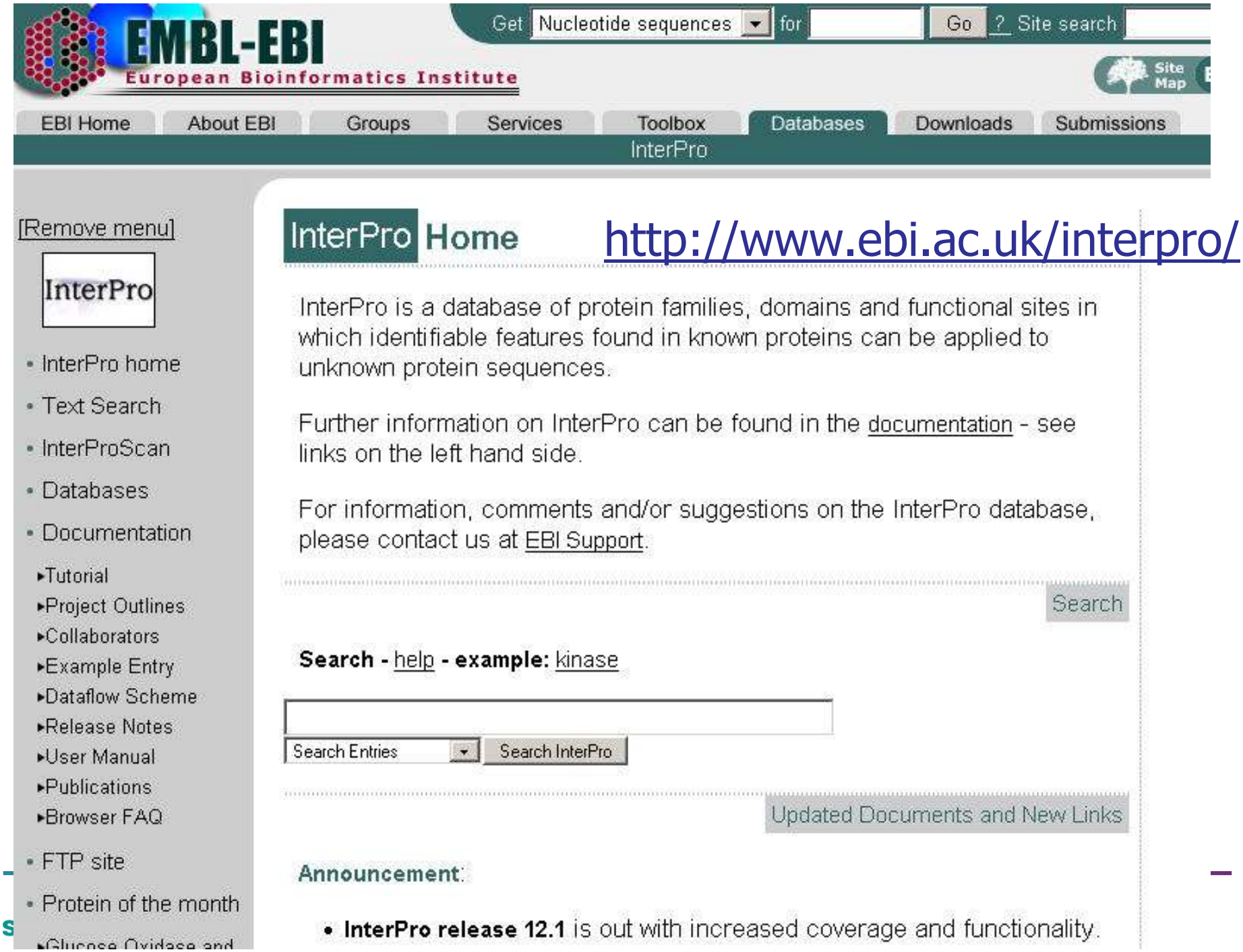
- Unification of PROSITE, PRINTS, Pfam into an integrated resource of protein families, domains and functional sites in 2000;
- Joint effort in creating a unified yet methodologically diverse system for protein family/domain identification;
- Single set of «documents» linked to the various methods;
- Distributed with tools by anonymous FTP and through WWW servers;
- Used to enhance the functional annotation of UniProtKB (Swiss-Prot and TrEMBL)
- progressively incorporates other databases

Current status of INTERPRO

Release 12.1 (March 2006) was built from Pfam, PRINTS, PROSITE, ProDom, SMART, TIGRFAMs, PIR SuperFamily, Scop based Superfamilies, Gene3D and PANTHER, and the current Swiss-Prot + TrEMBL data.
(for details see http://www.ebi.ac.uk/interpro/release_notes.html)

InterPro release 12.1 contains 12'953 entries, representing 3'585 domains, 9'055 families, 238 repeats, 32 active sites, 22 binding sites and 21 post-translational modification sites. Overall, there are 11'318'934 InterPro hits from 2'207'141 UniProtKB protein sequences.

92% of Swiss-Prot and 77% of TrEMBL protein sequences have one or more hits in InterPro.


EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search
Site Map
EBI Home
About EBI
Groups
Services
Toolbox
Databases
Downloads
Submissions
InterPro
[Remove menu]
InterPro
InterPro home
Text Search
InterProScan
Databases
Documentation
Tutorial
Project Outlines
Collaborators
Example Entry
Dataflow Scheme
Release Notes
User Manual
Publications
Browser FAQ
FTP site
Protein of the month
InterPro Home
http://www.ebi.ac.uk/interpro/
InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.
Further information on InterPro can be found in the documentation - see links on the left hand side.
For information, comments and/or suggestions on the InterPro database, please contact us at EBI Support.
Search
Search - help - example: kinase
Search Entries
Search InterPro
Updated Documents and New Links
Announcement:
InterPro release 12.1 is out with increased coverage and functionality.

http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001304

InterPro IPR001304 C-type lectin

Click here for help!

Matches

Overview: sorted by AC, sorted by name, of known structure, proteins with splice variants
Detailed: sorted by AC, sorted by name, of known structure proteins with splice variants
Table: For all matching proteins, of known structure
Architectures

Accession

IPR001304 Lectin_C Matches: 1925 proteins

Type

Domain

Signatures

Database	ID	Name	Proteins
Pfam	PF00059	Lectin_C	1749
PROSITE pattern	PS00615	C_TYPE_LECTIN_1	1119
PROSITE profile	PS50041	C_TYPE_LECTIN_2	1821
SMART	SM00034	CLECT	1777

Found in

IPR002352 Eosinophil major basic protein
IPR002353 Type II antifreeze protein
IPR002396 Selectin (CD62E/L/P antigen)
IPR003990 Pancreatitis-associated protein
IPR006228 Polycystin cation channel

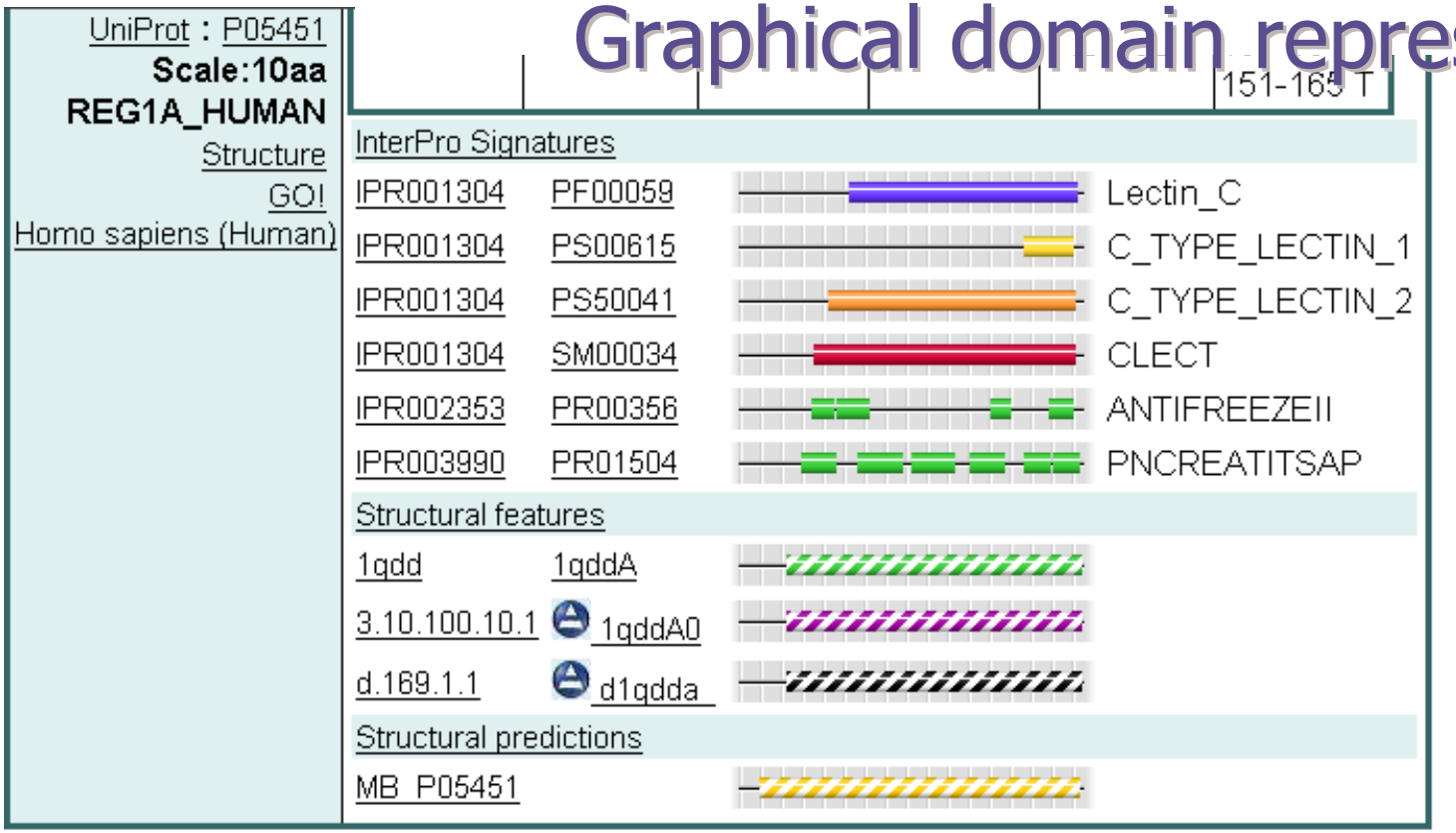
Function

GO:0005529 sugar binding

Animal lectins display a wide variety of architectures. They are classified according to the carbohydrate-recognition domain (CRD) of which there are two main types, S-type and C-type [1 , 2 , 3].

C-type lectins display a wide range of specificities. They require Ca^{2+} for their activity They are found predominantly but not exclusively in vertebrates.

InterPro: Graphical domain representation



ModBase	
PRINTS	
Pfam	
PROSITE profile	
PROSITE pattern	
SMART	
PDB Chain	
CATH Domain	
SCOP Domain	

Match Status key:	
T	True
F	False
P	Partial
U	Unknown
N	False negative

The ExPASy software tools

- Tools for the display and management of databases (NiceProt, Swiss-Shop sequence alerting system, etc.);
- Tools for sequence analysis (ScanProsite, ProtParam, ProtScale, RandSeq, Translate, etc.);
- Proteomics tools (AACompIdent, FindMod, FindPept, Aldente, PeptideMass, TagIdent, etc.);
- 3D-structure analysis and display tools (Swiss-Model, Swiss-PDBviewer)

ExPASy Proteomics tools

http://www.expasy.org/tools/

The tools marked by are local to the ExPASy server. The remaining tools are developed and hosted on other servers.

[Protein identification and characterization] [DNA -> Protein] [Similarity searches] [Pattern and profile searches] [Post-translational modifi
[Primary structure analysis] [Secondary structure prediction] [Tertiary structure] [Sequence alignment] [Phylogenetic analysis

Protein identification and characterization

Identification and characterization with peptide mass fingerprinting data

- Aldente - Identify proteins with peptide mass fingerprinting data. A new, fast and powerful tool that takes advantage of h recalibration and outlier exclusion
- FindMod - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and characterize the protein of interest.
- FindPept - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into ac post-translational modifications (PTM) and protease autolytic cleavage
- GlycoMod - Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses oligosaccharides and for glycopeptides)
- Mascot - Peptide mass fingerprint from Matrix Science Ltd., London
- PepMAPPER - Peptide mass fingerprinting tool from UMIST, UK
- PFMUTS - Shows the possible single and double mutations of a peptide fragment from MALDI peptide mass fingerprinting
- ProFound - Search known protein sequences with peptide mass information from Rockefeller and NY Universities [or from Gen
- ProteinProspector - UCSF tools for peptide masses data (MS-Fit, MS-Pattern, MS-Digest, etc.)

Identification and characterization with MS/MS data

- Popitam new - Identification and characterization tool for peptides with unexpected modifications (e.g. post-translational spectrometry
- Phenyx - Protein and peptide identification/characterization from MS/MS data from GeneBio, Switzerland
- Mascot - Sequence query and MS/MS ion search from Matrix Science Ltd., London
- OMSSA - MS/MS peptide spectra identification by searching libraries of known protein sequences
- PepFrag - Search known protein sequences with peptide fragment mass information from Rockefeller and NY Universities [or fi

- Use annotation in Swiss-Prot and TrEMBL (preprocessing, PTMs, etc.)
- Hyper-links between tools and databases

Identification:
Aldente,
TagIdent,
AAcompIdent,
MultiIdent

Characterization:
FindMod,
GlycoMod,
FindPept

Analysis:
PeptideMass,
GlycanMass,
BioGraph,
PeptideCutter
ProtScale,
ProtParam

Finding out about recent developments:

What's new on ExPASy:
http://www.expasy.org/history.html

UniProtKB/Swiss-Prot *recent* format changes:
http://www.expasy.org/sprot/relnotes/sp_news.html

UniProtKB/Swiss-Prot *planned* format changes:
http://www.expasy.org/sprot/relnotes/sp_soon.html

Subscribe to the electronic Swiss-Flash bulletins:
http://www.expasy.org/swiss-flash/

References

Swiss-Prot:

http://www.expasy.org/sprot/sprot-ref.html

Wu C. et al. *The Universal Protein Resource (UniProt): an expanding universe of protein information.*
Nucleic Acids Res. 34:D187-191(2006).

Boeckmann B. et al. *Protein variety and functional diversity: Swiss-Prot annotation in its biological context*
Comptes Rendus Biologies 328:882-99(2005).

Bairoch A.
Swiss-Prot: Juggling between evolution and stability
Brief. Bioinform. 5:39-55(2004).

Farriol-Mathis N. et al. *Annotation of post-translational modifications in the Swiss-Prot knowledgebase*. Proteomics 4:1537-1550(2004).

Gasteiger E. et al. A. *Swiss-Prot: Connecting biological knowledge via a protein database*
Curr. Issues Mol. Biol. 3:47-55(2001).

ExPASy:

Gasteiger E. et al. *ExPASy: the proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res. 31:3784-3788(2003).

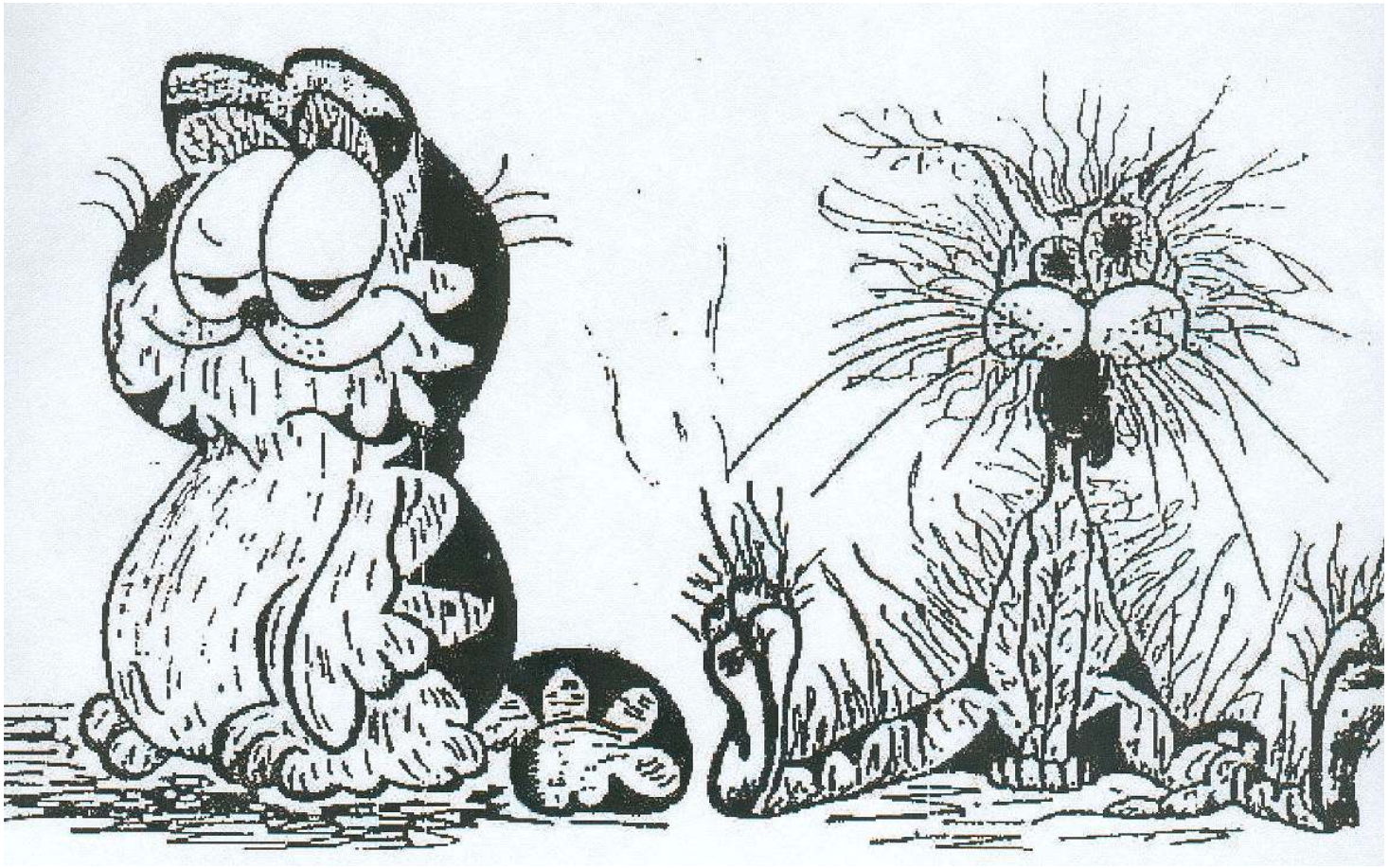
Useful general publications

- Nucleic Acids Res. Database issue 2006, vol. 34, supplement 1:

  http://nar.oupjournals.org/content/vol34/suppl_1/
- Nucleic Acids Res. Web server issue 2005, vol. 33, supplement 2:
  http://nar.oupjournals.org/content/vol33/suppl_2/
- Book: Bioinformatics for Dummies, by J.-M. Claverie and C. Notredame

  Publisher: For Dummies; 1st edition (January 15, 2003)
  ISBN: 0764516965

Before the introduction to Swiss-Prot…



After the introduction to Swiss-Prot…