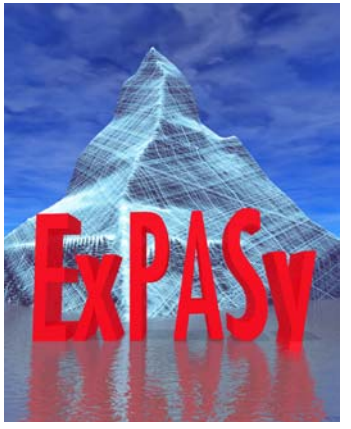


The challenges of extracting and representing protein information and knowledge

Amos Bairoch; Swiss Institute of Bioinformatics (SIB)

Swiss-Prot group

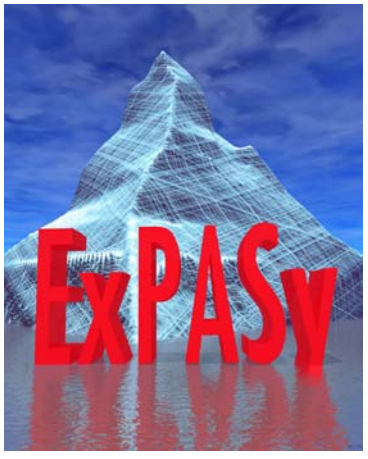
[BC]^2 Basel– March 18, 2004

Where we will be flying over in the next 30 minutes

1. Introduction;
2. The Anabelle annotation tool: how to extract more efficiently information from protein sequences using state of the art sequence analysis software;
3. How can we progress from the manual and time-consuming approach of extracting information from publications to a supervised text-mining driven mode;
4. NiceProt: a tool to explore the content of Swiss-Prot entries;
5. The representation of human disease mutations and polymorphisms at the level of the protein sequences in Swiss-Prot.

Swiss-Prot

- Created in July 1986; since 1987, a collaboration of the SIB and the EMBL/EBI; from 2003 onward it is the central part of the UniProt project;
- **Annotated**, **non-redundant**, **cross-referenced**, **documented** protein sequence **knowledge** resource;
- **145'000** sequences; 125'000 references; 1'300'000 cross-references; ~400 Mb of annotations;
- About **1'100'000** sequences in TrEMBL, the Swiss-Prot computer-annotated supplement;
- Weekly releases; available from about 50 servers, the main source being ExPASy.



The ExPASy WWW server

www.expasy.org

10 years of continuous service to the user community

- First molecular biology server on the Web (August 1993); ~350 million access since;
- Dedicated to proteomics:
  - Databases: **Swiss-Prot**, **PROSITE**, **Swiss-2DPAGE**, etc.;
  - Many 2D/MS protein identification/characterization and sequence analysis tools;
- Mirror sites in Australia, Bolivia, Canada, Korea, Taiwan and USA. Soon in Brazil.

UniProt
the universal protein resource

Text Search UniProt Knowledgebase

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

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

UniProt in one slide…

- Universal Protein Resource;
- Collaboration between 3 groups: the Swiss-Prot groups at SIB and EBI and the PIR group;
- www.uniprot.org (online since Dec 15, 2003);
- The UniProt Knowledgebase (UniProt) is the core component and is comprised of Swiss-Prot+TrEMBL;
- The UniProt Non-redundant Reference (UniRef) databases combine closely related sequences into a single record to speed searches. Three versions exist: UniRef50, UniRef90 and UniRef100;
- The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.

The contents of the Swiss-Prot protein knowledgebase

- Sequences!
- **ANNOTATIONS**
- References
- Taxonomic data
- Keywords
- Cross-references
- Documentation

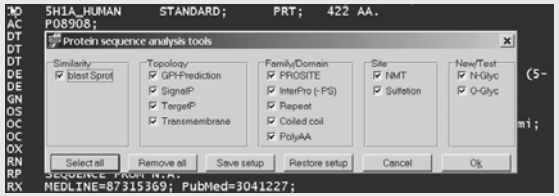
- Function(s); role(s)
- Post-translational modifications
- Domains
- Subcellular location
- Protein/protein interactions
- Similarities
- Diseases, mutagenesis
- Conflicts and variants

[2] Using sequence analysis to annotate protein sequences

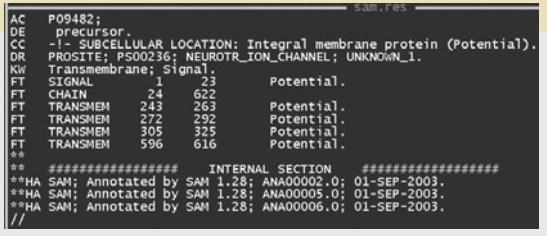
- Since the inception of Swiss-Prot, we have used a number of sequence analysis methods to help annotate protein sequences;
- Three periods:
  - Using PC/Gene (1986-1991);
  - Implementation of a number of tools, individually launched from our annotation platform (1992-2003);
  - Anabelle (2004-?).

Anabelle Integration and workflow

Swiss-Prot editor



macro



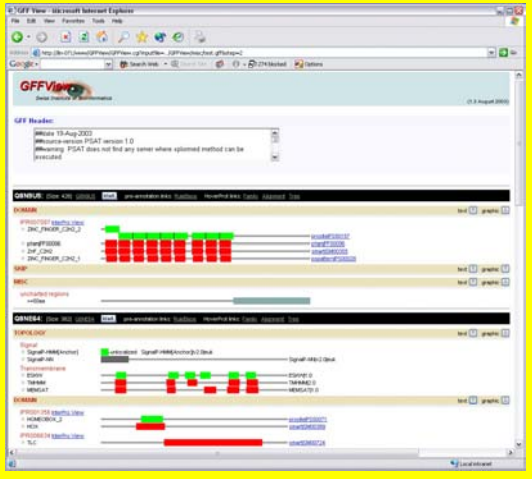
Anabelle

1) **PSAT**

Sequence Analysis programs

GFF data

2) **SelMo**



selected GFF data

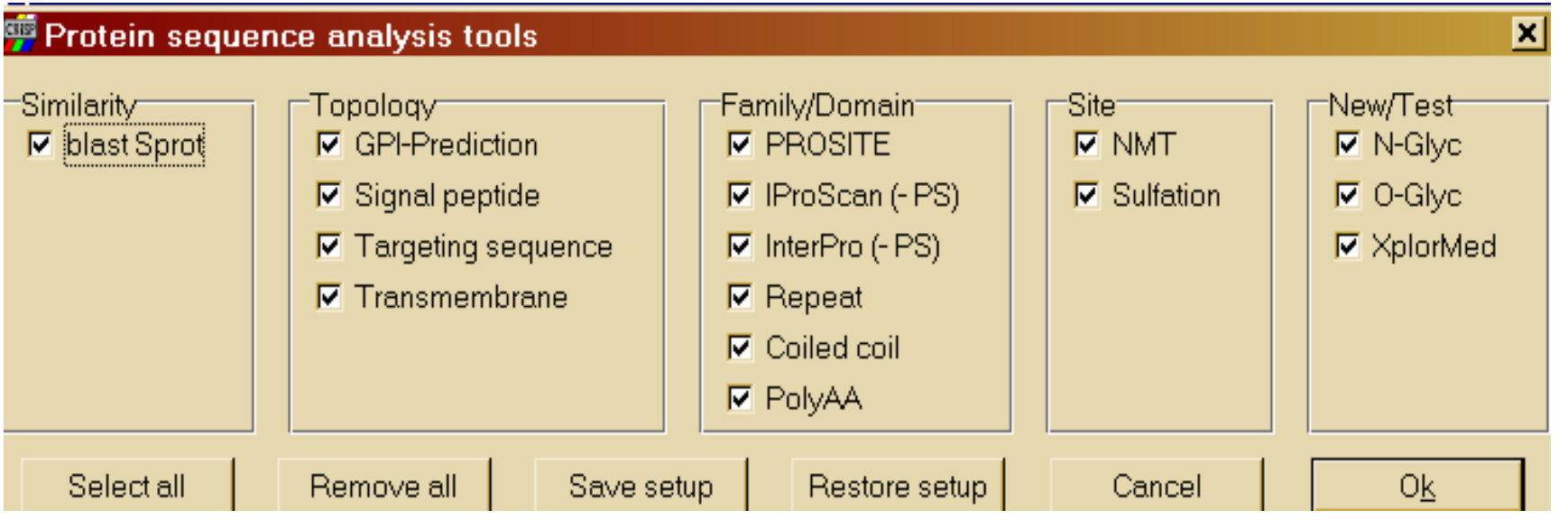
Swiss-Prot annotation

3) **SAM**

HAMAP

ProRules

PSAT



- PSAT runs many protein sequence analysis methods in parallel on one or several protein sequences and parses the results into the **general feature format (gff)**

- The input may be in Fasta or **Swiss-Prot** format.

Methods for the prediction of the protein topology

Methods	Results
SignalP	Prediction of a signal peptide by NN and HMM methods
TargetP	Chloroplast/Mitochondrion transit peptide
GPI_prediction	GPI-anchor prediction
Transmembrane	3 methods: TMHMM, MEMSAT, ESKW

Methods for the detection of protein families and domains

Methods	Results
ps_scan	PROSITE (profiles, patterns)
coils	Coiled-coils
REP	Repeats, e.g. Heat, Kelch, LRR, WD
polyAA	Streches of amino acids

Methods for the detection of post-translational modified amino acids

Methods	Results
Sulfinator	Prediction of potential tyrosine sulfation sites
NMT	Myristoylation of the N-terminal glycine (all) and internal glycine (viral sequence)
O-Glyc	O-glycosylation sites in mammalian proteins (Mucin-type GalNAc)
N-Glyc	N-Glycosylation sites in mammalian proteins

Results of PSAT are in the gff format

Example of a "general feature format" (gff) output:

INS_HUMAN SignalP-NN|v2.0|euk Signal 1 24|0.96 . . . Level 0 ; C-max "0.889,25,Y" ; Y-max "0.818,25,Y" ; S-max "0.998,7,Y" ; S-mean "0.902,Y" ; Category "TOPOLOGY"

- Defined fields

INS_HUMAN SignalP-NN|v2.0|euk Signal 1 24|0.96 . . .

ID	Swiss-Prot identifier
Source	Program name, the version number, parameters
Feature	Feature identifier, triggers the AA rule
From	Protein sequence position where the feature starts
To	Protein sequence position where the feature end
Score	Score for a hit

- Attributes

Level 0 ; C-max "0.889,25,Y" ; Y-max "0.818,25,Y" ; S-max "0.998,7,Y" ; S-mean "0.902,Y" ; Category "TOPOLOGY"

Level	Level of confidence in the result
Optional field	Values of results
Category	Method Class

SelMo

Viewer Layout:

GFF Header

for each protein:

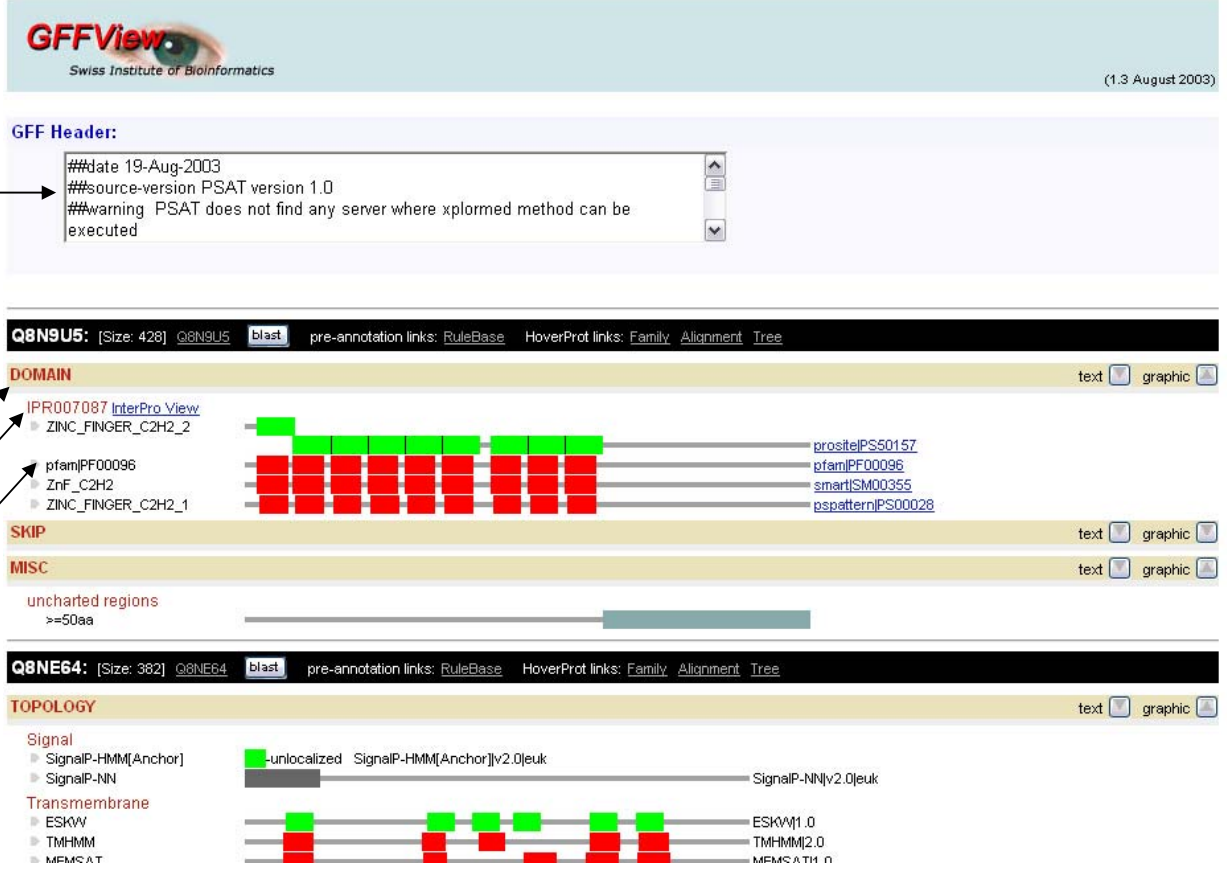
sort features by category (Topology, Domain, Best, Skip, Family, Similarity)

…by feature name

…by method name

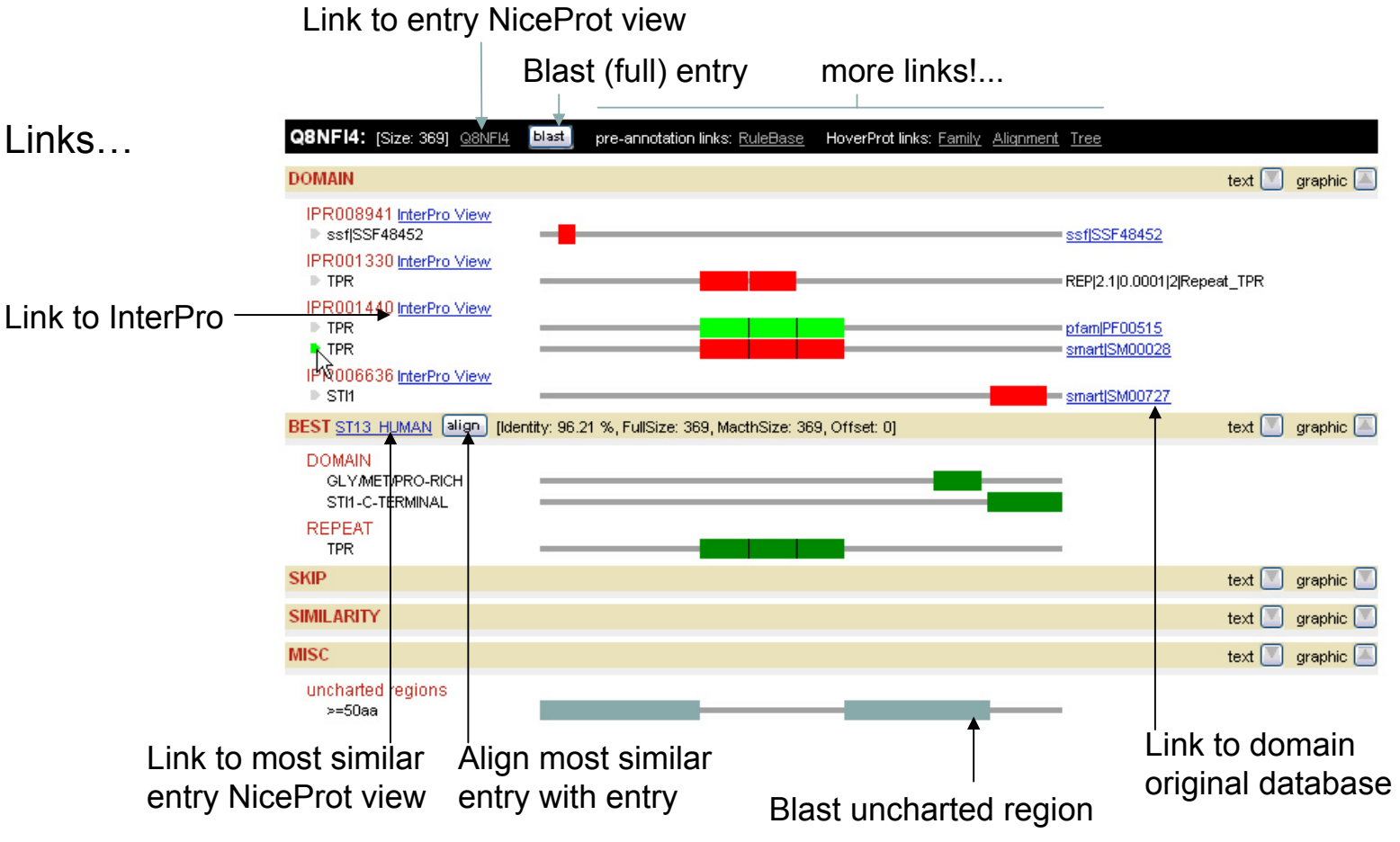
Selected feature

Unselected feature



SelMo

Viewer Layout:



```
CC   -!- SIMILARITY: Belongs to peptidase family S1.
CC   -!- SIMILARITY: Contains 2 CUB domains.
CC   -!- SIMILARITY: Contains 1 EGF-like domain.
CC   -!- SIMILARITY: Contains 2 Sushi (CCP/SCR) domains.
DR   PROSITE; PS00010; ASX_HYDROXYL; 1.
DR   PROSITE; PS01180; CUB; 2.
DR   PROSITE; PS01186; EGF_2; 1.
DR   PROSITE; PS01187; EGF_CA; 1.
DR   PROSITE; PS50923; SUSHI; 2.
DR   PROSITE; PS50240; TRYPSIN_DOM; 1.
DR   PROSITE; PS00134; TRYPSIN_HIS; 1.
DR   PROSITE; PS00135; TRYPSIN_SER; 1.
KW   Hydrolase; Serine protease; Protease; Sushi; Repeat; Signal;
KW   EGF-like domain; Hydroxylation.
FT   DOMAIN       20    138       CUB 1.
FT   DOMAIN      139    182       EGF-like, calcium-binding (Potential).
FT   DOMAIN      185    297       CUB 2.
FT   DOMAIN      299    364       Sushi 1.
FT   DOMAIN      365    434       Sushi 2.
FT   DOMAIN      449    699       Serine protease.
FT   ACT_SITE    490    490       Charge relay system (By similarity).
FT   ACT_SITE    552    552       Charge relay system (By similarity).
FT   ACT_SITE    646    646       Charge relay system (By similarity).
FT   MOD_RES     159    159       HYDROXYLATION (POTENTIAL).
FT   DISULFID     73     91       Potential.
FT   DISULFID    143    157       Potential.
FT   DISULFID    153    166       Potential.
FT   DISULFID    168    181       Potential.
FT   DISULFID    185    212       Potential.
FT   DISULFID    242    260       Potential.
FT   DISULFID    301    349       Potential.
FT   DISULFID    329    362       Potential.
FT   DISULFID    367    414       Potential.
FT   DISULFID    397    432       Potential.
FT   DISULFID    436    572       Potential.
FT   DISULFID    614    631       Potential.
FT   DISULFID    642    672       Potential.
FT   CARBOHYD     49     49       N-linked (GlcNAc...) (Potential).
FT   CARBOHYD    178    178       N-linked (GlcNAc...) (Potential).
FT   CARBOHYD    385    385       N-linked (GlcNAc...) (Potential).
FT   CARBOHYD    407    407       N-linked (GlcNAc...) (Potential).
```

The bio mint project

Anne-Lise Veuthey, Swiss Institute of Bioinformatics

SIB swissprot EBI

- 3 year FP5 European Project, started January 2003
- Official web site: www.biomint.org
- 5 teams involved:
  - University of Manchester (UK, coordinator)
  - PharmaDM (Belgium)
  - Austrian Research Institute for Artificial Intelligence (Austria)
  - University of Geneva, AI Lab (Switzerland)
  - University of Antwerp, CNTS (Belgium)
  - Swiss Institute of Bioinformatics (Switzerland)

The goals of BioMinT

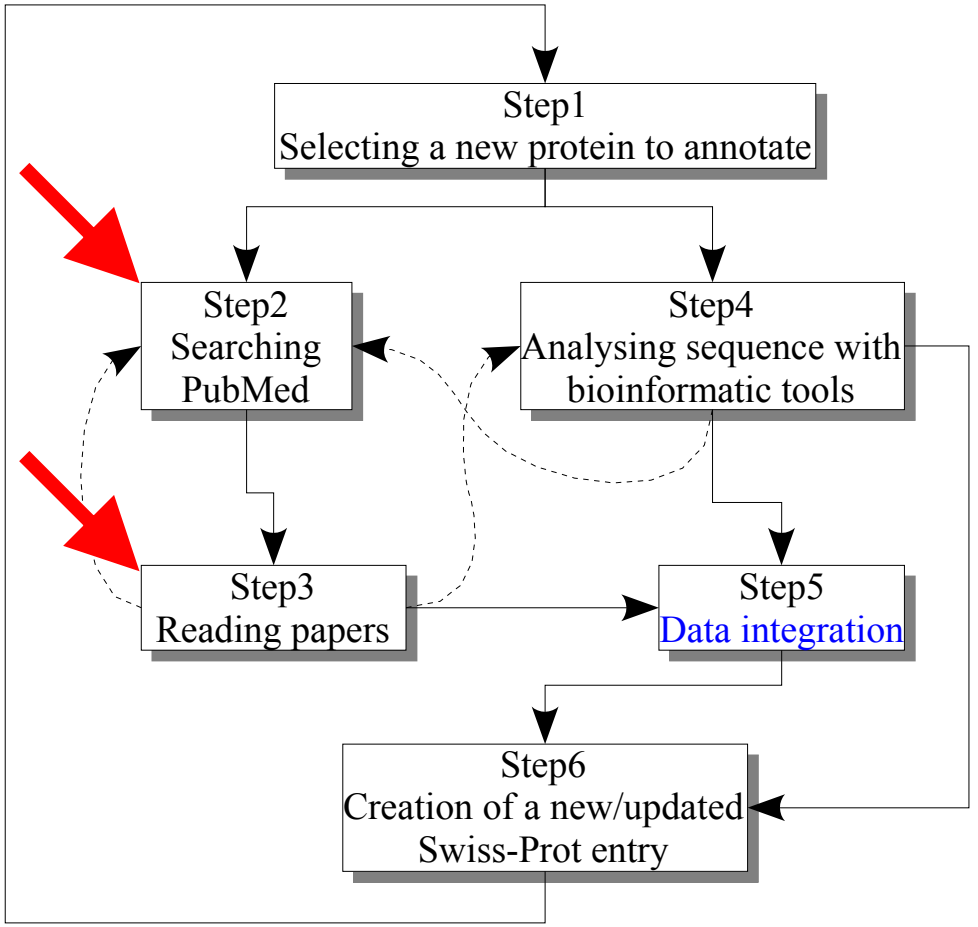
To develop a generic text mining tool that:

- Interprets different types of queries;
- Retrieves relevant documents from the biological literature;
- Extracts the required information;
- Outputs the result as a database slot filler or as a structured report

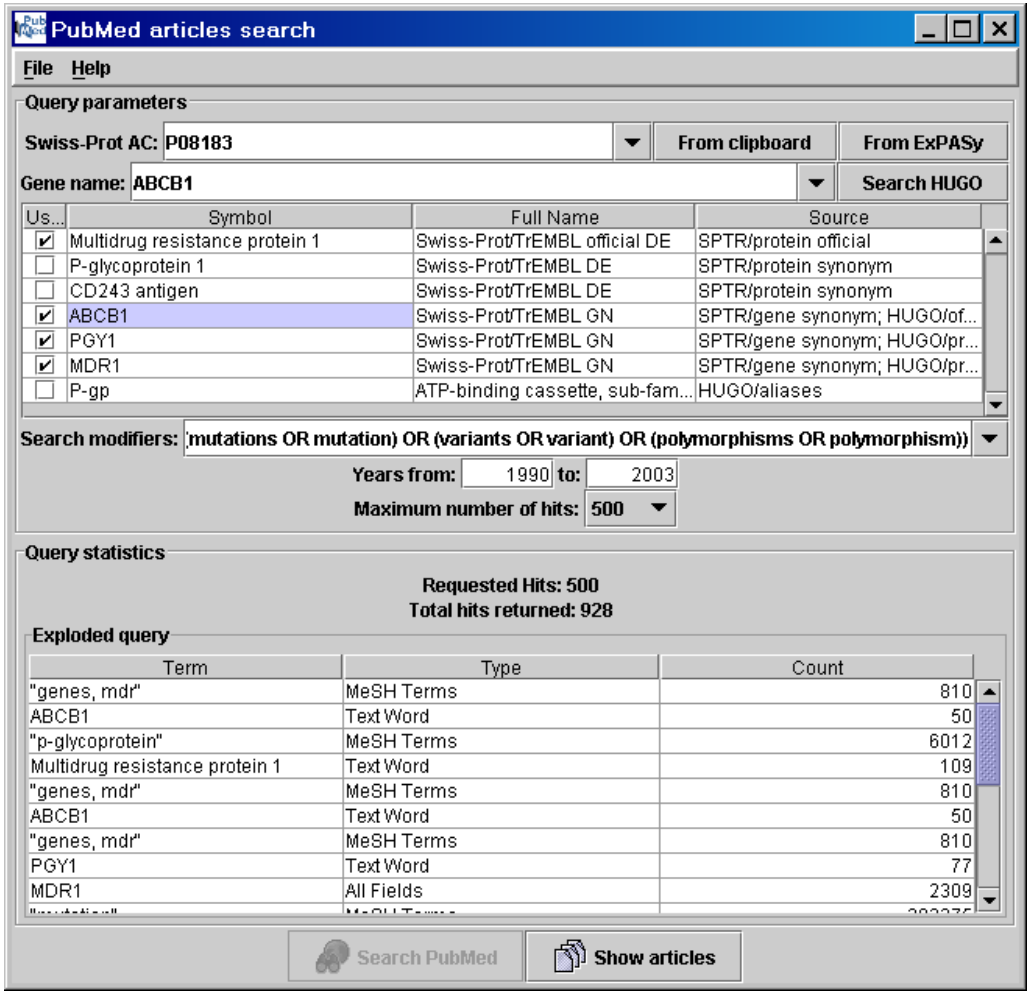
The tool will thus provide two essential research support services:

1. As a curator's assistant it will accelerate, by partially automating, the annotation and update of bio-databases;
2. As a researcher's assistant: it will generate readable reports in response to queries from biological researchers.

Swiss-Prot Entry Creation Flowchart



Query interface prototype



Developed by Pavel Dobrokhotov in the framework of SwissProt medical annotation: Bioinformatics 19(suppl. 1): i91-i94 (ISMB 2003)

PTM extractor

- A tool that relies on AA numbering to extract relevant features from PubMed abstracts;
- The tool makes use of the annotation in the target entry data to refine the positions (Example: taking into account a signal peptide);
- The current version is applied to the annotation of phosphorylation, N-glycosylation and palmitoylation sites, disulfide bridges, as well as to capture potential mutagenesis results.

Benchmark environment for training and evaluation

We need a corpus of supervised abstracts

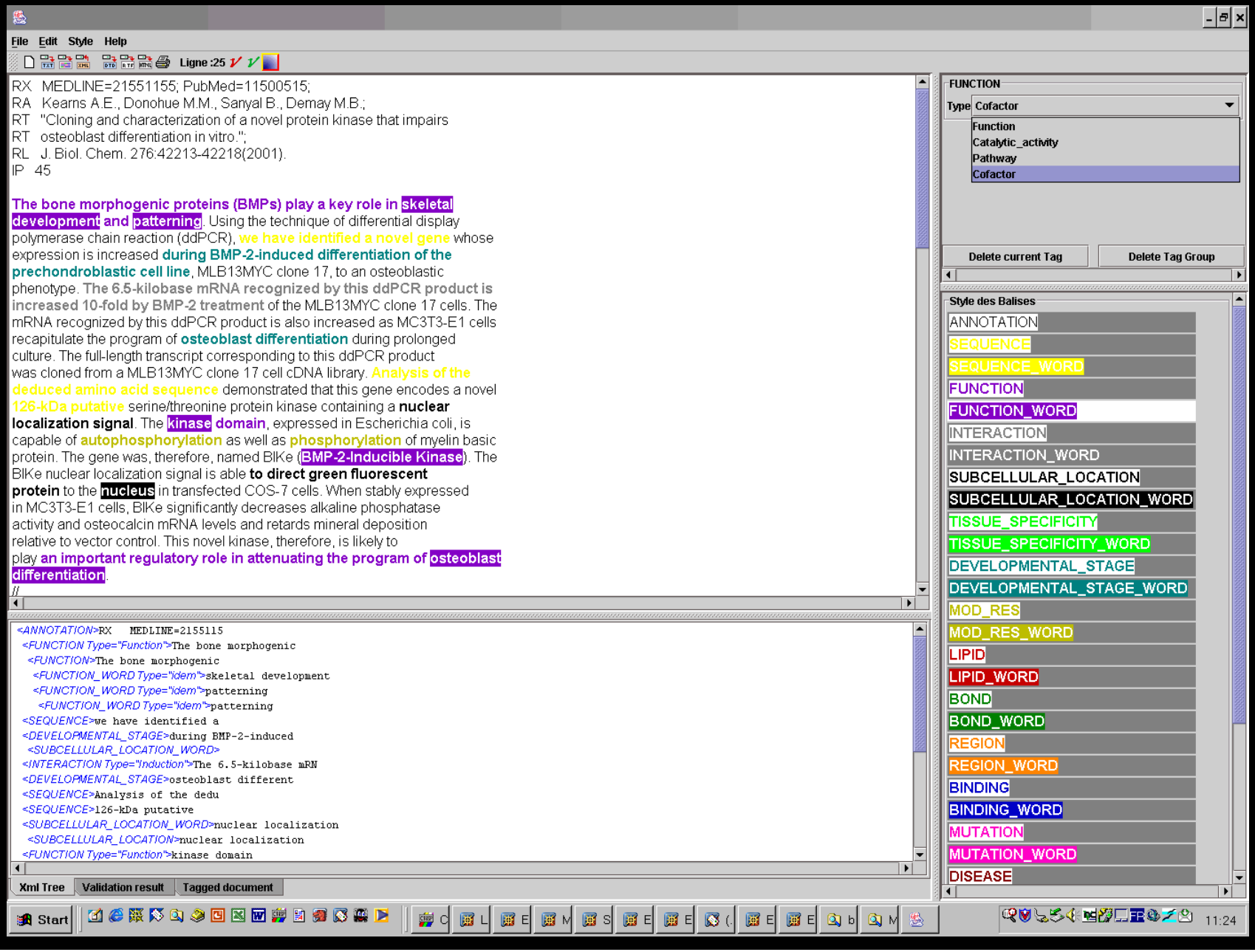
- ✓ To train the text-mining tools
- ✓ To elaborate rules for specific information extraction

What do we need to tag ?

- Fragments of, or whole sentences describing information useful for protein annotation
- Specific words describing a specific type of information

Document annotation interface

- Generic XML tagger currently being adapted for Swiss-Prot requirements;
- Developed by Gilles Bisson and Pierre-Emmanuel Gros, CNRS, Grenoble, in the framework of the Caderige project (http://caderige.imag.fr);
- Soon fully operational for use by the Swiss-Prot annotators.


File Edit Style Help
Ligne :25
RX MEDLINE=21551155; PubMed=11500515;
RA Kearns A.E., Donohue M.M., Sanyal B., Demay M.B.;
RT "Cloning and characterization of a novel protein kinase that impairs
RT osteoblast differentiation in vitro.";
RL J. Biol. Chem. 276:42213-42218(2001).
IP 45
The bone morphogenic proteins (BMPs) play a key role in skeletal development and patterning. Using the technique of differential display polymerase chain reaction (ddPCR), we have identified a novel gene whose expression is increased during BMP-2-induced differentiation of the prechondroblastic cell line, MLB13MYC clone 17, to an osteoblastic phenotype. The 6.5-kilobase mRNA recognized by this ddPCR product is increased 10-fold by BMP-2 treatment of the MLB13MYC clone 17 cells. The mRNA recognized by this ddPCR product is also increased as MC3T3-E1 cells recapitulate the program of osteoblast differentiation during prolonged culture. The full-length transcript corresponding to this ddPCR product was cloned from a MLB13MYC clone 17 cell cDNA library. Analysis of the deduced amino acid sequence demonstrated that this gene encodes a novel 126-kDa putative serine/threonine protein kinase containing a nuclear localization signal. The kinase domain, expressed in Escherichia coli, is capable of autophosphorylation as well as phosphorylation of myelin basic protein. The gene was, therefore, named BIKe (BMP-2-Inducible Kinase). The BIKe nuclear localization signal is able to direct green fluorescent protein to the nucleus in transfected COS-7 cells. When stably expressed in MC3T3-E1 cells, BIKe significantly decreases alkaline phosphatase activity and osteocalcin mRNA levels and retards mineral deposition relative to vector control. This novel kinase, therefore, is likely to play an important regulatory role in attenuating the program of osteoblast differentiation.
//
<ANNOTATION>RX MEDLINE=2155115
<FUNCTION Type="Function">The bone morphogenic
<FUNCTION>The bone morphogenic
<FUNCTION_WORD Type="idem">skeletal development
<FUNCTION_WORD Type="idem">patterning
<FUNCTION_WORD Type="idem">patterning
<SEQUENCE>we have identified a
<DEVELOPMENTAL_STAGE>during BMP-2-induced
<SUBCELLULAR_LOCATION_WORD>
<INTERACTION Type="Induction">The 6.5-kilobase mRN
<DEVELOPMENTAL_STAGE>osteoblast different
<SEQUENCE>Analysis of the dedu
<SEQUENCE>126-kDa putative
<SUBCELLULAR_LOCATION_WORD>nuclear localization
<SUBCELLULAR_LOCATION>nuclear localization
<FUNCTION Type="Function">kinase domain
Xml Tree
Validation result
Tagged document
FUNCTION
Type Cofactor
Function
Catalytic_activity
Pathway
Cofactor
Delete current Tag
Delete Tag Group
Style des Balises
ANNOTATION
SEQUENCE
SEQUENCE_WORD
FUNCTION
FUNCTION_WORD
INTERACTION
INTERACTION_WORD
SUBCELLULAR_LOCATION
SUBCELLULAR_LOCATION_WORD
TISSUE_SPECIFICITY
TISSUE_SPECIFICITY_WORD
DEVELOPMENTAL_STAGE
DEVELOPMENTAL_STAGE_WORD
MOD_RES
MOD_RES_WORD
LIPID
LIPID_WORD
BOND
BOND_WORD
REGION
REGION_WORD
BINDING
BINDING_WORD
MUTATION
MUTATION_WORD
DISEASE
Start
11:24

Function sentences

"Drosophila dishevelled (dsh) functions in two pathways: it is necessary to transduce Wingless (Wg) signaling and it is required in planar cell polarity."
<PMID=12072470>
[Drosophila dishevelled functions in at least 2 pathways - transduction of Wingless signalling and planar cell polarity]

"Following binding of a ligand to its cognate receptor, receptor-associated Jaks are activated. STAT proteins are then in turn activated by tyrosine phosphorylation by Jak kinases, allowing their dimerization and subsequent translocation into the nucleus, where they modulate expression of target genes."
<PMID=10781830>
[Activated Jaks phosphorylate tyrosines on STAT proteins, which allows them to dimerise, enter the nucleus and alter gene expression]

Disease sentences

"Tuberous sclerosis is a relatively common inherited disease that causes multiple benign tumours in different organs, frequently leading to skin rashes, seizures and mental handicap. The disease can be caused by mutations in either of two genes, TSC2, identified in 1993, and TSC1, only recently identified"
<PMID=9743993>
[Mutations in the TSC1 and TSC2 genes lead to tuberous sclerosis]

"Numerous studies have demonstrated that chemokines play an integral role in diseases marked by inflammation"
<PMID=12476351>
[Chemokines are involved in diseases marked by inflammation]

NiceProt View of Swiss-Prot:
P00750

Printer-friendly view | Submit 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	**TPA_HUMAN**
Primary accession number	**P00750**
Secondary accession number	Q15103
Entered in Swiss-Prot in	Release 01, July 1986
Sequence was last modified in	Release 01, July 1986
Annotations were last modified in	Release 43, March 2004
Name and origin of the protein	
Protein name	**Tissue-type plasminogen activator [Precursor]**
Synonyms	**EC 3.4.21.68** **tPA** **t-PA** **t-plasminogen activator** **Alteplase** **Reteplase**
Gene name	**PLAT**
From	Homo sapiens (Human) [TaxID: 9606]
Taxonomy	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.

References

[1] SEQUENCE FROM NUCLEIC ACID.
TISSUE=Melanoma;
MEDLINE=83115262; PubMed=6337343; [NCBI, ExPASy, EBI, Israel, Japan]
Pennica D., Holmes W.E., Kohr W.J., Harkins R.N., Vehar G.A., Ward C.A., Bennett W.F., Yelverton E., Seeburg P.H., Heyneker H.L., Goeddel D.V., Collen D.;
"Cloning and expression of human tissue-type plasminogen activator cDNA in E. coli.";
Nature 301:214-221(1983).

[2] SEQUENCE FROM NUCLEIC ACID.
TISSUE=Fetal lung;
MEDLINE=88262579; PubMed=3133640; [NCBI, ExPASy, EBI, Israel, Japan]
Sasaki H., Saito Y., Hayashi M., Otsuka K., Niwa M.;
"Nucleotide sequence of the tissue-type plasminogen activator cDNA from human fetal lung cells.";
Nucleic Acids Res. 16:5695-5695(1988).

[3] SEQUENCE FROM NUCLEIC ACID.
MEDLINE=88054470; PubMed=2824147; [NCBI, ExPASy, EBI, Israel, Japan]
Reddy V.B., Garramone A.J., Sasak H., Wei C.-M., Watkins P., Galli J., Hsiung N.;
"Expression of human uterine tissue-type plasminogen activator in mouse cells using BPV vectors.";
DNA 6:461-472(1987).

[22] STRUCTURE BY NMR OF KRINGLE 2.
MEDLINE=92106329; PubMed=1762144; [NCBI, ExPASy, EBI, Israel, Japan]
Byeon I.-J.L., Llinas M.;
"Solution structure of the tissue-type plasminogen activator kringle 2 domain complexed to 6-aminohexanoic acid an antifibrinolytic drug.";
J. Mol. Biol. 222:1035-1051(1991).

[23] STRUCTURE BY NMR OF 38-85.
MEDLINE=92292163; PubMed=1602484; [NCBI, ExPASy, EBI, Israel, Japan]
Downing A.K., Driscoll P.C., Harvey T.S., Dudgeon T.J., Smith B.O., Baron M., Campbell I.D.;
"Solution structure of the fibrin binding finger domain of tissue-type plasminogen activator determined by 1H nuclear magnetic resonance.";
J. Mol. Biol. 225:821-833(1992).

[24] STRUCTURE BY NMR OF 36-126.
MEDLINE=96027104; PubMed=7582899; [NCBI, ExPASy, EBI, Israel, Japan]
Smith B.O., Downing A.K., Driscoll P.C., Dudgeon T.J., Campbell I.D.;
"The solution structure and backbone dynamics of the fibronectin type I and epidermal growth factor-like pair of modules of tissue-type plasminogen activator.";
Structure 3:823-833(1995).

Comments

- ***FUNCTION***: Converts the abundant, but inactive, zymogen plasminogen to plasmin by hydrolyzing a single Arg-Val bond in plasminogen. By controlling plasmin-mediated proteolysis, it plays an important role in tissue remodeling and degradation, in cell migration and many other physiopathological events.
- ***CATALYTIC ACTIVITY***: Specific cleavage of Arg-|-Val bond in plasminogen to form plasmin.
- ***SUBUNIT***: Heterodimer of chain A and chain B held by a disulfide bond. Binds to fibrin with high affinity. This interaction leads to an increase in the catalytic efficiency of the enzyme between 100-and 1000-fold, due to an increase in affinity for plasminogen. Similarly, binding to heparin increases the activation of plasminogen. Binding to laminin and fibronectin has also been demonstrated. Binds to mannose receptor and the low-density lipoprotein receptor-related protein (LRP1). These proteins are involved in TPA clearance. Also binds to annexin II and to cytokeratin 8. Yet unidentified interactions on endothelial cells and vascular smooth muscle cells (VSMC) lead to a 100-fold stimulation of plasminogen activation. In addition, binding to VSMC reduces TPA inhibition by PAI-1 by 30-fold.
- ***SUBCELLULAR LOCATION***: Secreted; extracellular.
- ***ALTERNATIVE PRODUCTS***:
  - **Alternative splicing [2 named forms]** Display all isoform sequences in Fasta format

Name	Long
Isoform ID	P00750-1
This is the isoform sequence displayed in this entry.	

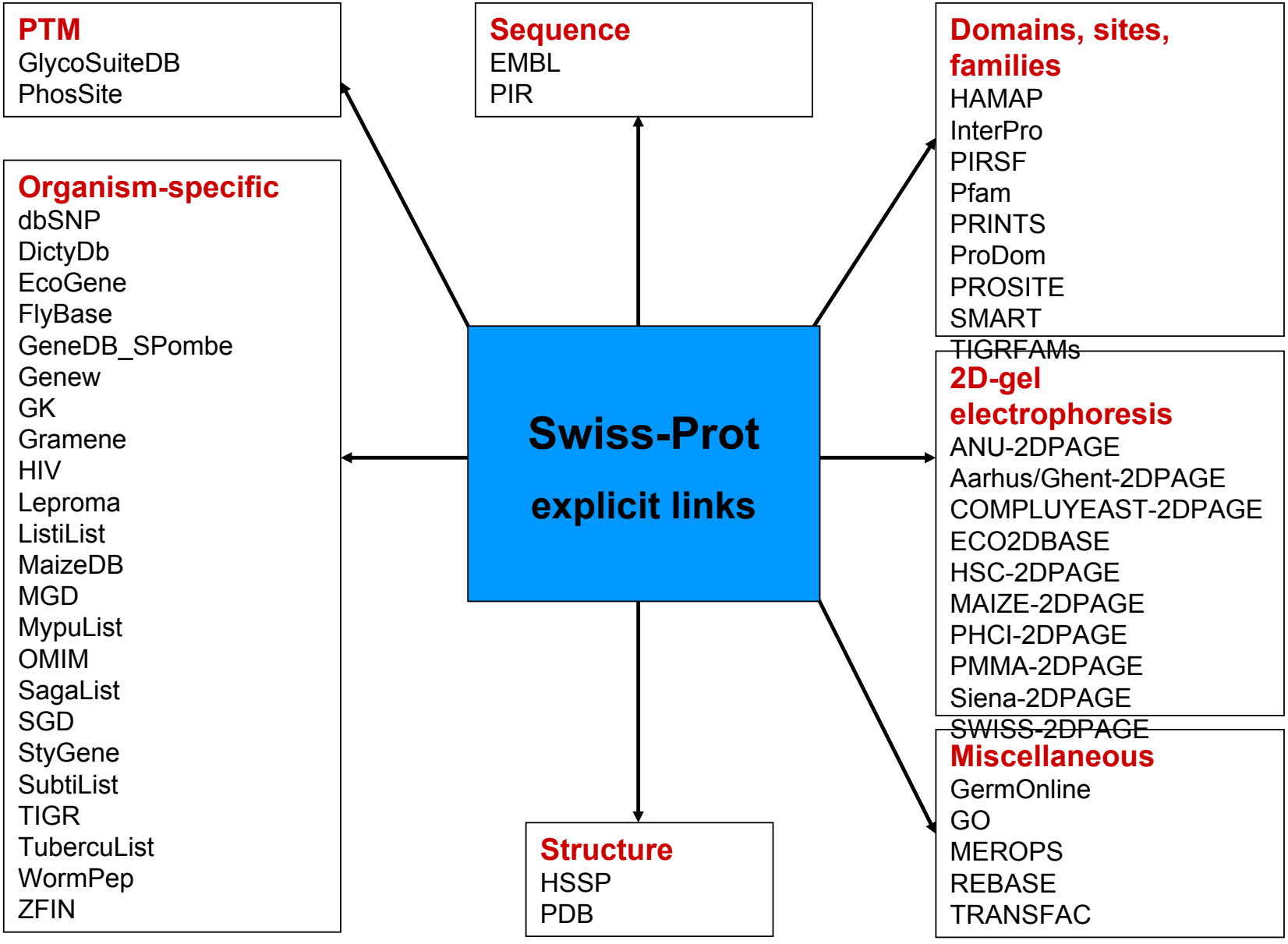
Name	Short
Isoform ID	P00750-2
Note: No experimental confirmation available.	
Features which should be applied to build the isoform sequence: VSP_005411, VSP_005412.	

- ***TISSUE SPECIFICITY***: Synthesized in numerous tissues (including tumors) and secreted into most extracellular body fluids, such as plasma, uterine fluid, saliva, gingival crevicular fluid, tears, seminal fluid, milk.
- ***DOMAIN***: Both FN1 and one of the kringle domains are required for binding to fibrin.
- ***DOMAIN***: Both FN1 and EGF-like domains are important for binding to LRP1.
- ***DOMAIN***: The FN1 domain mediates binding to annexin II.
- ***DOMAIN***: The second kringle domain is implicated in binding to cytokeratin 8 and to the endothelial cell surface binding site.
- ***PTM***: The single chain, almost fully active enzyme, can be further processed into a two-chain fully active form by a cleavage after Arg-310 catalyzed by plasmin, tissue kallikrein or factor Xa.

- ***PTM***: Differential cell-specific N-linked glycosylation gives rise to two variants: type I and type II. The single chain type I variant is less readily converted into the two-chain form by plasmin and the two-chain type I variant has a lower activity than the two-chain type II in the presence of fibrin.
- ***PTM***: N-glycosylation of Asn-152; the bound oligomannosidic glycan is involved in the interaction with the mannose receptor.
- ***PTM***: Characterization of O-linked glycan was studied in Bowes melanoma cell line.
- ***DISEASE***: Increased activity of TPA causes hyperfibrinolysis, with excessive bleeding as a consequence.
- ***DISEASE***: Defective release of TPA causes hypofibrinolysis, leading to thrombosis or embolism.
- ***PHARMACEUTICAL***: Available under the names Activase (Genentech) and Retavase (Centocor and Roche) [Retavase is a fragment of TPA that contains kringle 2 and the protease domain; it was also known as BM 06.022]. Used in Acute Myocardial Infarction (AMI), in Acute Ischemic Stroke (AIS) and Pulmonary Embolism (PE) to initiates fibrinolysis.
- ***SIMILARITY***: Belongs to peptidase family S1.
- ***SIMILARITY***: Contains 1 EGF-like domain.
- ***SIMILARITY***: Contains 1 fibronectin type I domain.
- ***SIMILARITY***: Contains 2 kringle domains.
- ***DATABASE***: NAME=Activase; NOTE=Clinical information on Activase; WWW="http://www.genentech.com/gene/products/information/cardiovascular/activase/".
- ***DATABASE***: NAME=Retavase; NOTE=Clinical information on Retavase; WWW="http://www.centocor.com/cgi-bin/site/products/prod_retavase.cgi".

Cross-references

Cross-references	
EMBL	L00153; AAB59510.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] L00141; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00142; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00143; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00144; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00145; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00146; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00147; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00148; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00149; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00150; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] L00151; AAB59510.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] X07393; CAA30302.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] A07197; CAA00642.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M18182; AAA36800.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] K03021; AAA98809.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M15518; AAA60111.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] X13097; CAA31489.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] BC007231; AAH07231.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] V00570; CAA23833.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M11890; AAA61213.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] M11889; AAA61213.1; JOINED. [EMBL / GenBank / DDBJ] [CoDingSequence] D01096; BAA00881.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence] A01465; CAA00166.1; -. [EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A94004; UKHUT.
PDB	1BDA; 11-MAY-99. [ExPASy / RCSB / EBI] 1PK2; 31-JAN-94. [ExPASy / RCSB / EBI] 1TPG; 15-SEP-95. [ExPASy / RCSB / EBI] 1TPK; 15-JUL-93. [ExPASy / RCSB / EBI] 1TPM; 31-JAN-94. [ExPASy / RCSB / EBI] 1TPN; 31-JAN-94. [ExPASy / RCSB / EBI] 1RTF; 11-JAN-97. [ExPASy / RCSB / EBI] 1A5H; 20-APR-99. [ExPASy / RCSB / EBI] 1PML; 22-JUN-94. [ExPASy / RCSB / EBI] Detailed list of linked structures.


PTM
GlycoSuiteDB
PhosSite
Sequence
EMBL
PIR
Domains, sites, families
HAMAP
InterPro
PIRSF
Pfam
PRINTS
ProDom
PROSITE
SMART
TIGRFAMs
Organism-specific
dbSNP
DictyDb
EcoGene
FlyBase
GeneDB_SPombe
Genew
GK
Gramene
HIV
Leproma
ListiList
MaizeDB
MGD
MypuList
OMIM
SagaList
SGD
StyGene
SubtiList
TIGR
TubercuList
WormPep
ZFIN
Swiss-Prot
explicit links
2D-gel electrophoresis
ANU-2DPAGE
Aarhus/Ghent-2DPAGE
COMPLUYEAST-2DPAGE
ECO2DBASE
HSC-2DPAGE
MAIZE-2DPAGE
PHCI-2DPAGE
PMMA-2DPAGE
Siena-2DPAGE
SWISS-2DPAGE
Miscellaneous
GermOnline
GO
MEROPS
REBASE
TRANSFAC
Structure
HSSP
PDB

MEROPS	S01.232; -.
GlycoSuiteDB	P00750; -.
Genew	HGNC:9051; PLAT.
CleanEx	HGNC:9051; PLAT.
GeneCards	PLAT.
GeneLynx	PLAT; Homo sapiens.
GenAtlas	PLAT.
MIM	173370 [NCBI / EBI].
GO	GO:0004296; Molecular function: t-plasminogen activator activity *(traceable author statement)*. GO:0007596; Biological process: blood coagulation *(traceable author statement)*. GO:0006464; Biological process: protein modification *(traceable author statement)*. GO:0006508; Biological process: proteolysis and peptidolysis *(traceable author statement)*.
SOURCE	PLAT; Homo sapiens.
Ensembl	P00750; Homo sapiens. [Entry / Contig view]
InterPro	IPR009003; Cys_Ser_trypsin. IPR006209; EGF_like. IPR000083; Fibrnctn1. IPR006210; IEGF. IPR000001; Kringle. IPR001254; Peptidase_S1. IPR001314; Peptidase_S1A. Graphical view of domain structure.
Pfam	PF00008; EGF; 1. PF00039; fn1; 1. PF00051; kringle; 2. PF00089; trypsin; 1. Pfam graphical view of domain structure.
PRINTS	PR00722; CHYMOTRYPSIN. PR00018; KRINGLE.
ProDom	PD000395; Kringle; 2. [Domain structure / List of seq. sharing at least 1 domain]

Cross-references can be explicit or implicit and can also be a source of data

SMART	SM00181; EGF; 1. SM00058; FN1; 1. SM00130; KR; 2. SM00020; Tryp_SPc; 1.
PROSITE	PS00022; EGF_1; 1. PS01186; EGF_2; 1. PS50026; EGF_3; 1. PS01253; FIBRONECTIN_1; 1. PS00021; KRINGLE_1; 2. PS50070; KRINGLE_2; 2. PS50240; TRYPSIN_DOM; 1. PS00134; TRYPSIN_HIS; 1. PS00135; TRYPSIN_SER; 1.
HOVERGEN	[Family / Alignment / Tree]
BLOCKS	P00750.
ProtoNet	P00750.
ProtoMap	P00750.
PRESAGE	P00750.
DIP	P00750.
ModBase	P00750.
SMR	P00750; B7EC9B1A5E3FDC4D.
SWISS-2DPAGE	Get region on 2D PAGE.

Keywords

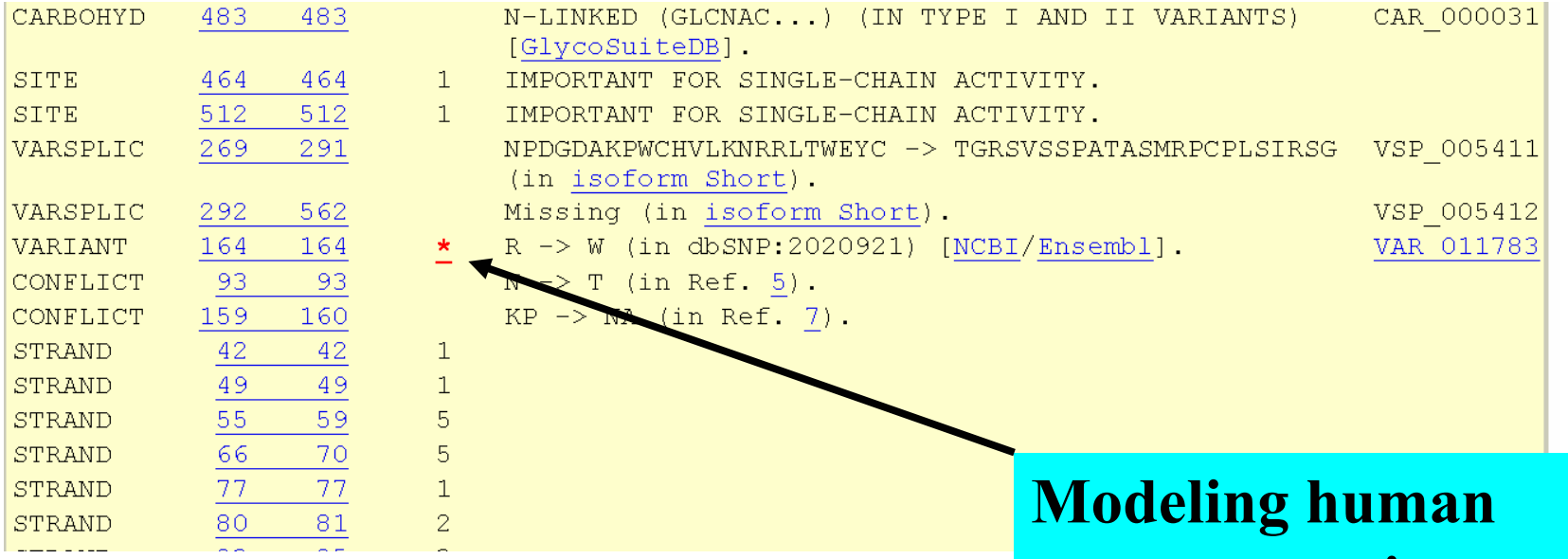
Plasminogen activation; **Hydrolase**; **Serine protease**; **Glycoprotein**; **Plasma**; **Kringle**; **EGF-like domain**; **Repeat**; **Signal**; **Alternative splicing**; **Pharmaceutical**; **3D-structure**; **Polymorphism**.

Key	From	To	Length	Description	FTId
SIGNAL	1	23	23	*Potential*.	
PROPEP	24	32	9		
PROPEP	33	35	3	Removed by plasmin.	
CHAIN	36	562	527	TISSUE-TYPE PLASMINOGEN ACTIVATOR.	
CHAIN	36	310	275	TISSUE-TYPE PLASMINOGEN ACTIVATOR A CHAIN.	
CHAIN	311	562	252	TISSUE-TYPE PLASMINOGEN ACTIVATOR B CHAIN.	
DOMAIN	39	81	43	FIBRONECTIN TYPE-I.	
DOMAIN	82	120	39	EGF-LIKE.	
DOMAIN	127	208	82	KRINGLE 1.	
DOMAIN	215	296	82	KRINGLE 2.	
DOMAIN	311	562	252	SERINE PROTEASE.	
DOMAIN	42	52	11	IMPORTANT FOR BINDING TO ANNEXIN II.	
SITE	102	102	1	IMPORTANT FOR BINDING TO LRP1.	
ACT_SITE	357	357		CHARGE RELAY SYSTEM.	
ACT_SITE	406	406		CHARGE RELAY SYSTEM.	
ACT_SITE	513	513		CHARGE RELAY SYSTEM.	
DISULFID	41	71			
DISULFID	69	78			
DISULFID	86	97			
DISULFID	91	108			
DISULFID	110	119			
DISULFID	127	208		*By similarity*.	
DISULFID	148	190		*By similarity*.	
DISULFID	179	203		*By similarity*.	
DISULFID	215	296			
DISULFID	236	278			
DISULFID	267	291			
DISULFID	299	430		INTERCHAIN.	
DISULFID	342	358		*By similarity*.	
DISULFID	350	419		*By similarity*.	
DISULFID	444	519		*By similarity*.	
DISULFID	476	492		*By similarity*.	
DISULFID	509	537		*By similarity*.	
CARBOHYD	96	96		O-LINKED (FUC) [GlycoSuiteDB].	CAR_000029
CARBOHYD	152	152		N-LINKED (GLCNAC...) (IN TYPE I AND II VARIANTS).	
CARBOHYD	219	219		N-LINKED (GLCNAC...) (IN TYPE I VARIANT ONLY) [GlycoSuiteDB].	CAR_000030
CARBOHYD	483	483		N-LINKED (GLCNAC...) (IN TYPE I AND II VARIANTS) [GlycoSuiteDB].	CAR_000031

Many different features:

Domains

PTMs



CARBOHYD	483	483		N-LINKED (GLCNAC...) (IN TYPE I AND II VARIANTS) [GlycoSuiteDB].	CAR_000031
SITE	464	464	1	IMPORTANT FOR SINGLE-CHAIN ACTIVITY.	
SITE	512	512	1	IMPORTANT FOR SINGLE-CHAIN ACTIVITY.	
VARSPLIC	269	291		NPDGDAKPWCHVLKNRRLTWEYC -> TGRSVSSPATASMRPCPLSIRSG (in isoform Short).	VSP_005411
VARSPLIC	292	562		Missing (in isoform Short).	VSP_005412
VARIANT	164	164	*	R -> W (in dbSNP:2020921) [NCBI/Ensembl].	VAR_011783
CONFLICT	93	93		N -> T (in Ref. 5).	
CONFLICT	159	160		KP -> NA (in Ref. 7).	
STRAND	42	42	1		
STRAND	49	49	1		
STRAND	55	59	5		
STRAND	66	70	5		
STRAND	77	77	1		
STRAND	80	81	2		

Modeling human sequence variants

Features, continued

Representing variations: variants, splice variants, mutagenesis and sequence conflicts

HELIX 553 559 7

Sequence information

Length: **562 AA** [This is the length of the unprocessed precursor]	Molecular weight: **62917 Da** [This is the MW of the unprocessed precursor]	CRC64: **B7EC9B1A5E3FDC4D** [T on the sequence]

```
        10         20         30         40         50         60
         |          |          |          |          |          |
MDAMKRGLCC VLLLCGAVFV SPSQEIHARF RRGARSYQVI CRDEKTQMIY QQHQSWLRPV

        70         80         90        100        110        120
         |          |          |          |          |          |
LRSNRVEYCW CNSGRAQCHS VPVKSCSEPR CFNGGTCQQA LYFSDFVCQC PEGFAGKCCE

       130        140        150        160        170        180
         |          |          |          |          |          |
IDTRATCYED QGISYRGTWS TAESGAECTN WNSSALAQKP YSGRRPDAIR LGLGNHNYCR

       190        200        210        220        230        240
         |          |          |          |          |          |
NPDRDSKPWC YVFKAGKYSS EFCSTPACSE GNSDCYFGNG SAYRGTHSLT ESGASCLPWN

       250        260        270        280        290        300
         |          |          |          |          |          |
SMILIGKVYT AQNPSAQALG LGKHNYCRNP DGDAKPWCHV LKNRRLTWEY CDVPSCSTCG

       310        320        330        340        350        360
         |          |          |          |          |          |
LRQYSQPQFR IKGGLFADIA SHPWQAAIFA KHRRSPGERF LCGGILISSC WILSAAHCFQ

       370        380        390        400        410        420
         |          |          |          |          |          |
ERFPPHHLTV ILGRTYRVVP GEEEQKFEVE KYIVHKEFDD DTYDNDIALL QLKSDSSRCA

       430        440        450        460        470        480
         |          |          |          |          |          |
QESSVVRTVC LPPADLQLPD WTECELSGYG KHEALSPFYS ERLKEAHVRL YPSSRCTSQH
```

The Human Proteomics Initiative

HPI

- Annotation of all known human proteins;
- Annotation of mammalian orthologs of human proteins;
- Annotation of all known human polymorphisms at the protein sequence level;
- Annotation of all known post-translational modifications in human proteins;
- Tight links to structural information.

Current state of human protein information in Swiss-Prot

- 10’600 annotated human sequences;
- Associated with about 27’000 distinct literature references;
- 25’000 experimental or predicted post-translational modifications;
- 4’000 splice variants;
- 18’000 polymorphisms (most of which are linked with disease states).

Protein Polymorphisms

- Called '**c-SNPs**' (coding single nucleotide polymorphisms) or '**SAPs**' (single amino-acid polymorphisms);
- Mutations that cause major changes to a protein sequence (such as frameshift mutations) are not considered to be relevant to Swiss-Prot, as their deleterious effect on a given protein's function is usually obvious.

```
FT   VARIANT     117    117       A -> V (linked to development of
FT                                dementing Gerstmann-Straussler disease).
FT                                /FTId=VAR_006466.
FT   VARIANT     129    129       M -> V (polymorphism; determines the
FT                                disease phenotype in patients who have a
FT                                PrP mutation at position 178. Patients
FT                                with M-129 develop FFI, those with V-129
FT                                develop CJD; dbSNP:1799990).
FT                                /FTId=VAR_006467.
FT   VARIANT     131    131       G -> V (in GSD).
FT                                /FTId=VAR_014264.
FT   VARIANT     171    171       N -> S (in schizoaffective disorder).
FT                                /FTId=VAR_006468.
FT   VARIANT     178    178       D -> N (in FFI and CJD).
FT                                /FTId=VAR_006469.
FT   VARIANT     180    180       V -> I (in CJD).
FT                                /FTId=VAR_006470.
FT   VARIANT     183    183       T -> A (in familial spongiform
FT                                encephalopathy).
FT                                /FTId=VAR_006471.
```

Swiss-Prot variant: VAR_009543 in P53634

[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.

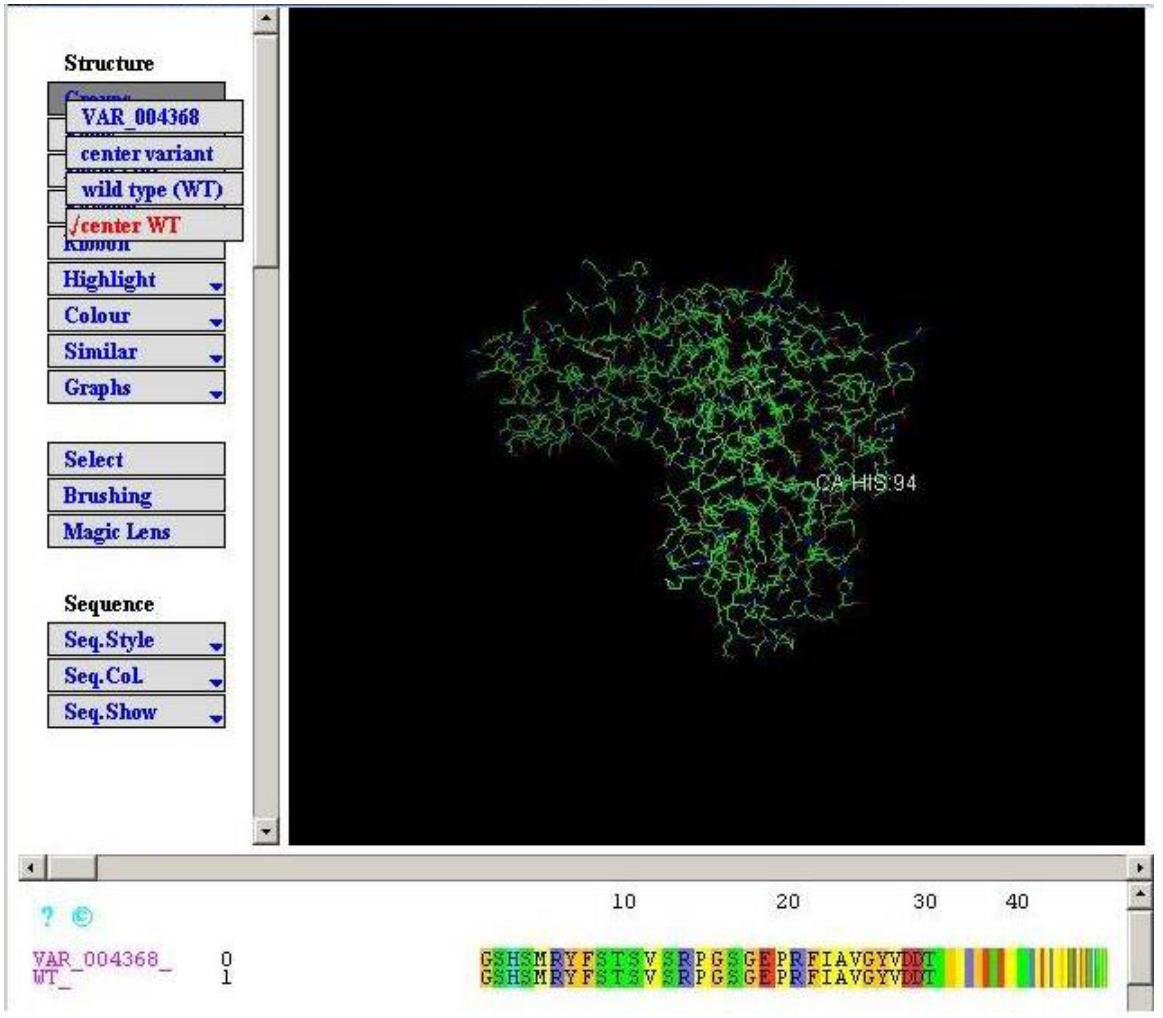
General information	
Swiss-Prot ID (AC)	**CATC_HUMAN (P53634)**
Gene symbol(s)	**Official:** CTSC **Synonym(s):** CPPI
Chromosomal location	11q14.1-q14.3
Genew	HGNC: 2528
Protein name	Dipeptidyl-peptidase I [Precursor]
Length of the protein	463
Information on the variant	
FTId	**VAR_009543**
Amino acid position of the variant	272
Residue change	**From Arg (R) to Pro (P), R272P**
Status (Disease, polymorphism or unclassified)	Disease
Disease	**Papillon-Lefevre syndrome (PLS)** Defects in CTSC are a cause of Papillon-Lefevre syndrome (PLS) [MIM:245000]; also known as keratosis palmoplantaris with periodontopathia. PLS is an autosomal recessive disorder that is mainly ascertained by dentists because of the severe periodontitis that afflicts patients. Both the deciduous and permanent dentitions are affected, resulting in premature tooth loss. Palmoplantar keratosis, varying from mild psoriasiform scaly skin to overt hyperkeratosis, typically develops within the first three years of life. Keratosis also affects other sites such as elbows and knees
Comment	None
Structural information on the variant	
Location on the sequence	252 QASCGSCYSFASMGMLEARI R ILTNNSQTPILSPQEVVSCS 292 ? P
Protein features in neighborhood	**Key From To Length Description** CHAIN 231 394 164 DIPEPTIDYL-PEPTIDASE I BETA CHAIN
Residue conservation	Alignment from Blast search
Physico-chemical property	Change from large size and basic (R) to medium size and hydrophobic (P)
3D homology models	**Model Visualization**: ExPASy, AstexViewer **Template Structure**: 1K3BB [ExPASy / EBI-MSD] ***Disclaimer:*** The result of any modelling procedure is non-experimental and must be considered with care. This is especially true since there is no human intervention during model building process.

References for the variant
[1] VARIANTS PLS PHE-249; LEU-252; PRO-272; SER-301; CYS-339 AND CYS-347. MEDLINE=20047769; PubMed=10581027; [NCBI, ExPASy, EBI, Israel, Japan] Toomes C., James J., Wood A.J., Wu C.L., McCormick D., Lench N., Hewitt C., Moynihan L., Roberts E., Woods C.G., Markham A., Wong M., Widmer R., Ghaffar K.A., Pemberton M., Hussein I.R., Temtamy S.A., Davies R., Read A.P., Sloan P., Dixon M.J., Thakker N.S. "Loss-of-function mutations in the cathepsin C gene result in periodontal disease and palmoplantar keratosis." Nat. Genet. 23:421-424(1999)
[2] VARIANTS PLS PRO-127; PRO-272; CYS-339 AND CYS-429, AND VARIANTS ILE-153 AND LYS-401. MEDLINE=21884204; PubMed=11886537; [NCBI, ExPASy, EBI, Israel, Japan] Lefevre C., Blanchet-Bardon C., Jobard F., Bouadjar B., Stalder J.-F., Cure S., Hoffmann A., Prud'Homme J.-F., Fischer J. "Novel point mutations, deletions, and polymorphisms in the cathepsin C gene in nine families from Europe and North Africa with Papillon-Lefevre syndrome." J. Invest. Dermatol. 117:1657-1661(2001)

Cross-references for the variant	
OMIM	245000 [NCBI / EBI].
dbSNP	Not available
HGVbase	Not available
Specialized database	Not available

For more information on the content of this page, please have a look at the documentation.

Visualisation of 3D models



Target → Template Selection

Template selection

Swiss-Prot human sequences

Search: Blastp

Chainsaw

Selection:

- Crystal structure
- ≥70 % sequence Identity
- Resolution ≤ 2.5 Å
- No side-chain atom error

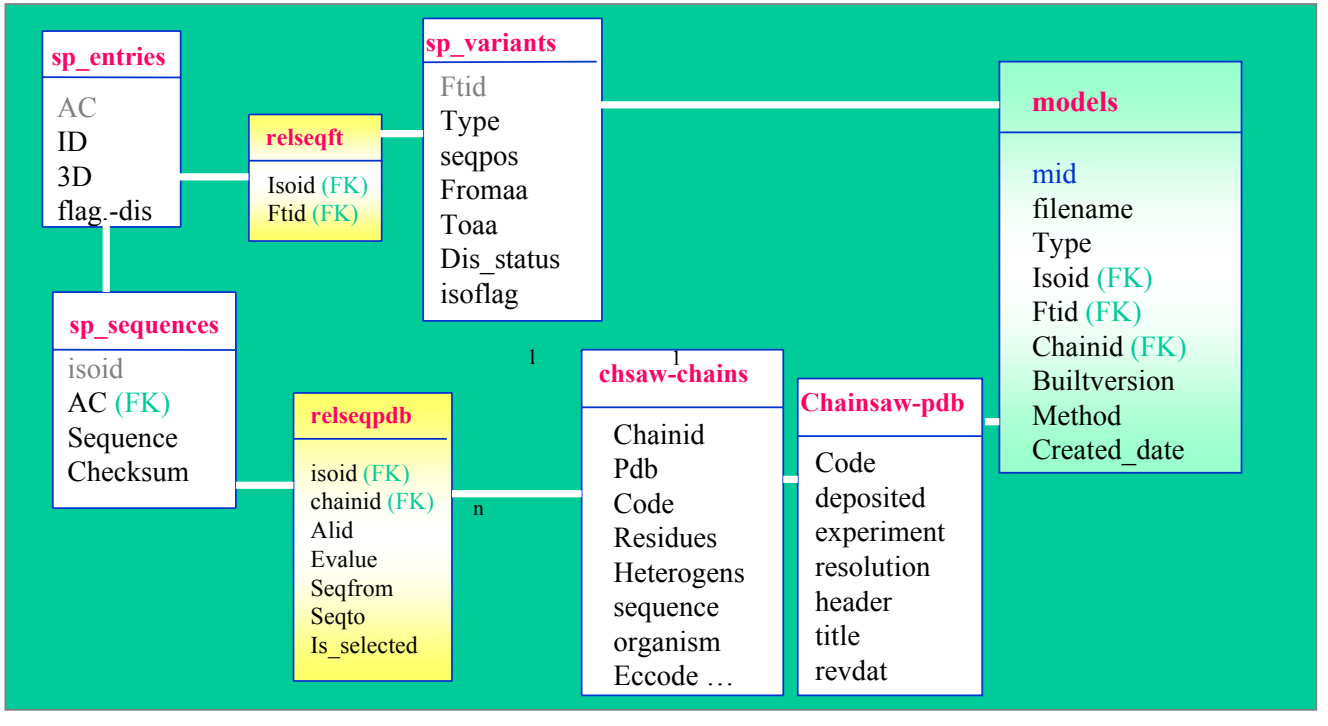
Template superposition

Discard Redundant templates

Final template list

The ModSNP database

- Routine update
  - According to PDB, Swiss-Prot entries
- Automatic generation of high-quality homology models



ModSNP: a basis for a large-scale structural analysis of protein variants

- Main advantages:
  - Large number of models;
  - Each variants are classified into disease-related, polymorphism, or unclassified according to Swiss-Prot annotation;
  - Routine automatic update

The Swiss-Prot staff at SIB and EBI

- **Group leaders:** Amos Bairoch, Rolf Apweiler
- **Annotators/curators:** Philippe Aldebert, Andrea Auchincloss, Ghislaine Argoud Puy, Kristian Axelsen, Kirsty Bates, Margaret Biswas, Marie-Claude Blatter, Brigitte Boeckmann, Silvia Braconi Quintaje, Paul Browne, Evelyn Camon, Danielle Coral, Elisabeth Coudert, Tania de Oliveira Lima, Kirill Degtyarenko, Sylvie Dethiollaz, Ruth Eberhardt, Anne Estreicher, Livia Famiglietti, Nathalie Farriol-Mathis, Serenella Ferro, Gill Fraser, John Garavelli, Raffaella Gatto, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz-Gumowski, Ursula Hinz, Chantal Hulo, Nicolas Hulo, Janet James, Florence Jungo, Vivien Junker, Youla Karavidopoulou, Guillaume Keller, Maria Krestyaninova, Kati Laiho, Petra Langendijk-Genevaux, Minna Lehvaslaiho, David Lonsdale, Michele Magrane, Karine Michoud, Virginie Mittard, Madelaine Moinat, Nicola Mulder, Claire O'Donovan, Sandra Orchard, Sandrine Pilbout, Sylvain Poux, Manuela Prüss, Sorogini Reynaud, Catherine Rivoire, Bernd Röchert, Michel Schneider, Christian Sigrist, André Stutz, Shyamala Sundaram, Michael Tognoli, Elmar von Baum, Sandra van den Broek, Eleanor Whitfield
- **Programmers and system administrators:** Daniel Barrell, David Binns, Laurent Bollondi, Sergio Contrino, Michael Darsow, Edouard deCastro, Séverine Duvaud, Stephanie Federico, Alexander Fedetov, Astrid Fleischmann, Wolfgang Fleischmann, Elisabeth Gasteiger, Alain Gateau, Andre Hackmann, Henning Hermjakob, Ivan Ivanyi, Eric Jain , Alexander Kanapin, Paul Kersey, Ernst Kretschmann, Corinne Lachaize, Maria-Jesus Martin, Xavier Martin, John O 'Rourke, Tom Oinn, Isabelle Phan, Astrid Rakow, Nicole Redaschi, Kai Runte, Florence Servant, Allyson Williams, Dan Wu
- **Research staff:** Pavel Dobrokhotov, Alexandre Gattiker, Margus Luk, Anne Morgat, Violaine Pillet, Anne-Lise Veuthey, Lina Yip
- **Clerical and secretarial assistance:** Veronique Mangold, Claudia Sapsezian, Margaret Shore-Nye, Laure Verbregue