Protein databases

Henrik Nielsen

Protein databases, historical background

Swiss-Prot, http://www.expasy.org/sprot/

Established in 1986 in Switzerland

ExPASy (Expert Protein Analysis System)

Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

PIR, http://pir.georgetown.edu/

Established in 1984

National Biomedical Research Foundation, Georgetown University, USA

In 2002 merged into:

UniProt, http://www.uniprot.org/

A collaboration between SIB, EBI and Georgetown University.

UniProt

UniProt Knowledgebase (UniProtKB)

UniProt Reference Clusters (UniRef)

UniProt Archive (UniParc)

UniProt Knowledgebase Release 2022_03 (03-Aug-2022) consists of:

UniProtKB/Swiss-Prot: Annotated manually (*curated*)

568,002 entries

UniProtKB/TrEMBL: Computer annotated

226,771,948 entries

Types of databases

GenBank / EMBL / DDBJ:

- Entries created & maintained by individual contributors
- No check for redundancy

Swiss-Prot:

- Entries created & maintained by staff
- Better standards compliance

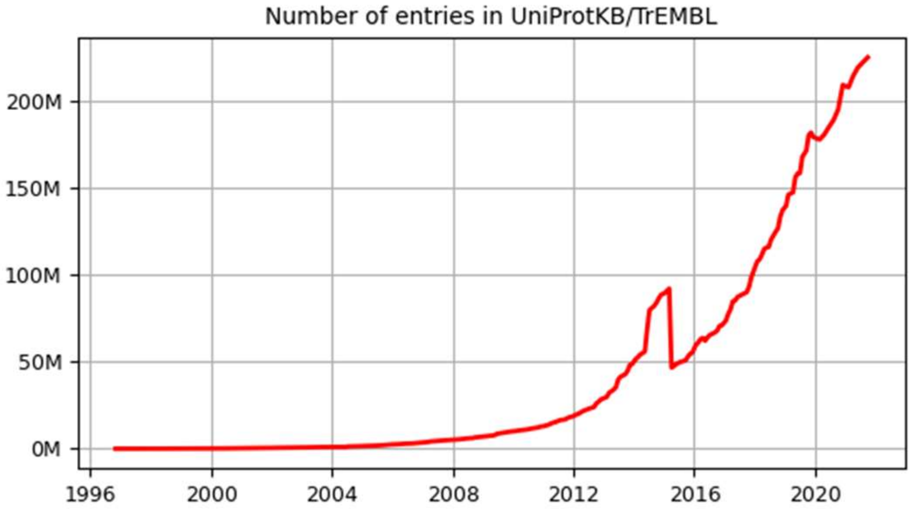
TrEMBL:

- Entries created by automatic translation of EMBL sequences & annotations

Growth of UniProt

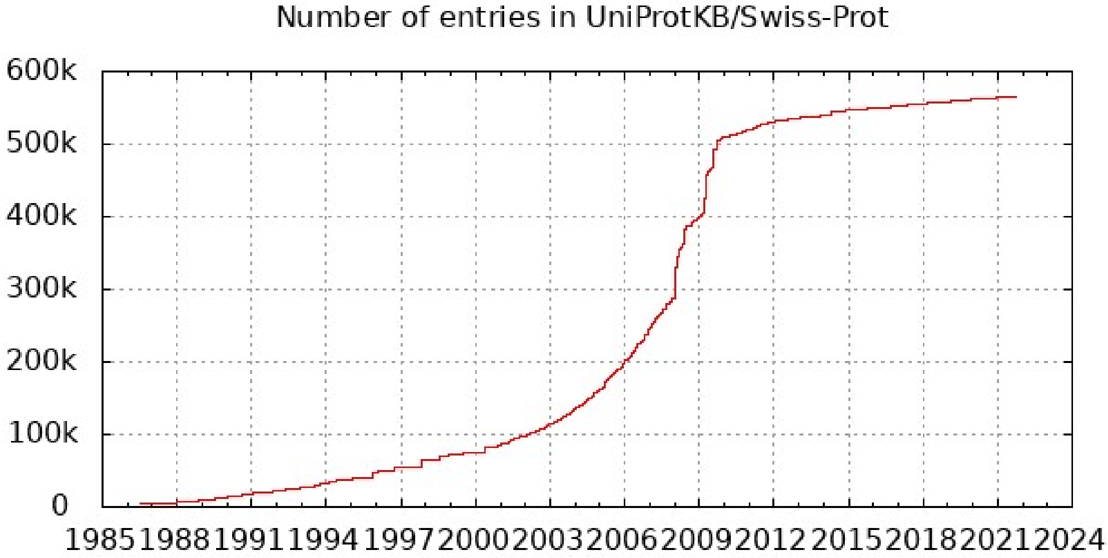
TrEMBL

https://www.ebi.ac.uk/uniprot/TrEMBLstats



Swiss-Prot

https://web.expasy.org/docs/relnotes/relstat.html



Content of UniProt Knowledgebase

- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

Amino acid sequences

From where do you get amino acid sequences?

- Translation of nucleotide sequences (GenBank/EMBL/DDBJ)
- Direct amino acid sequencing: *Edman degradation*
- Mass spectrometry
- 3D-structures

Content of UniProt Knowledgebase

- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

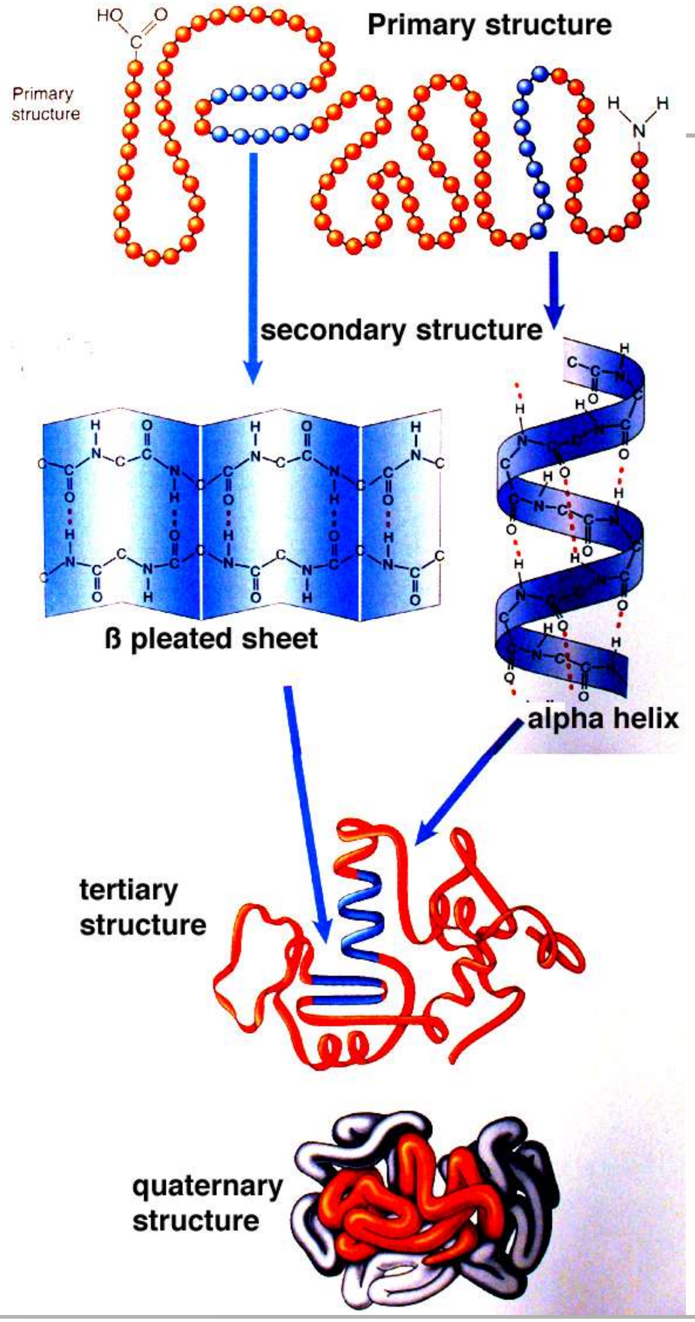
Protein structure

Primary structure: Amino acid sequence

Secondary structure:
”Backbone” hydrogen bonding
Alpha helix / Beta sheet / Turn

Tertiary structure: Fold, 3D coordinates

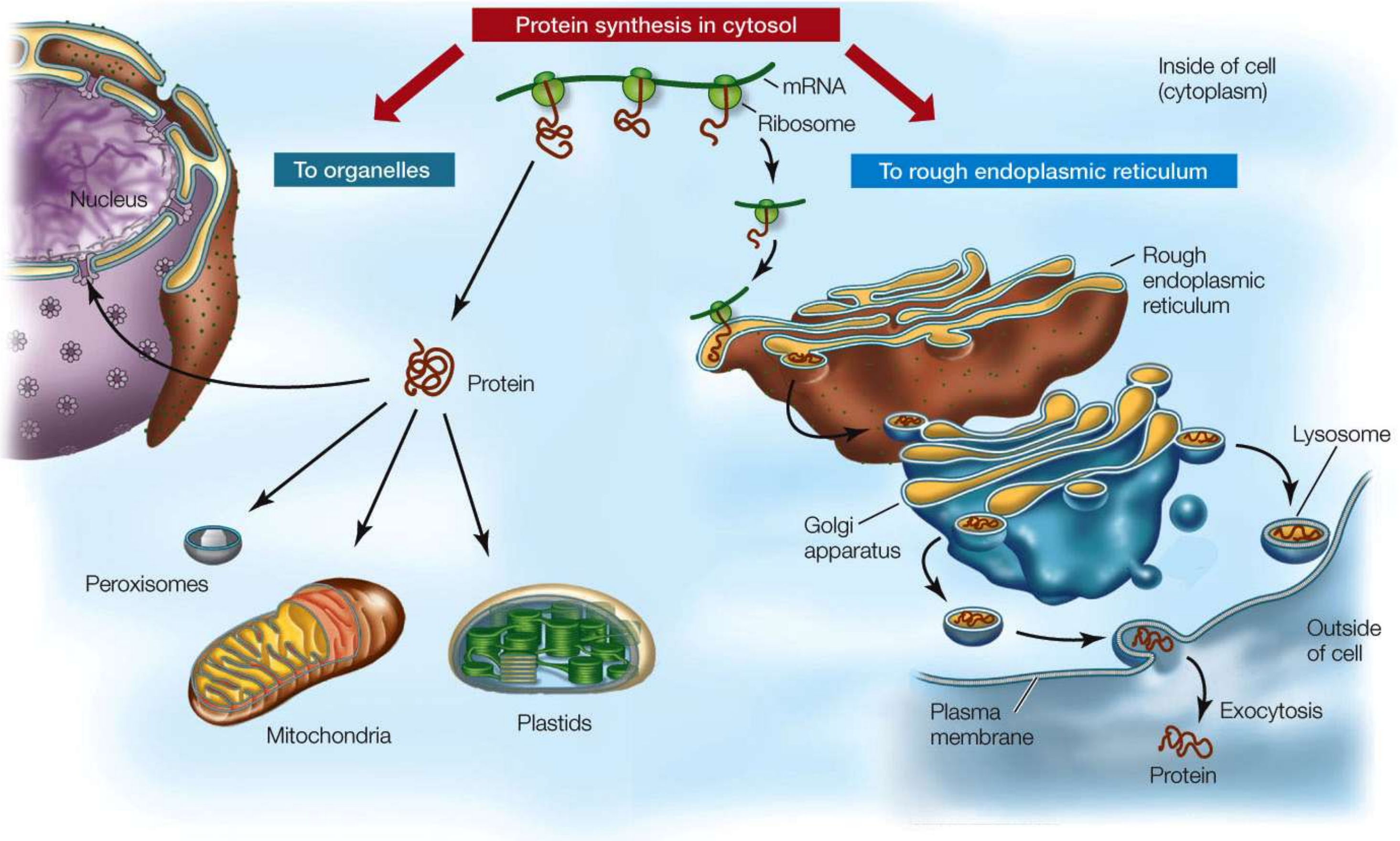
Quaternary structure: subunits



Content of UniProt Knowledgebase

- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

Subcellular location / protein sorting

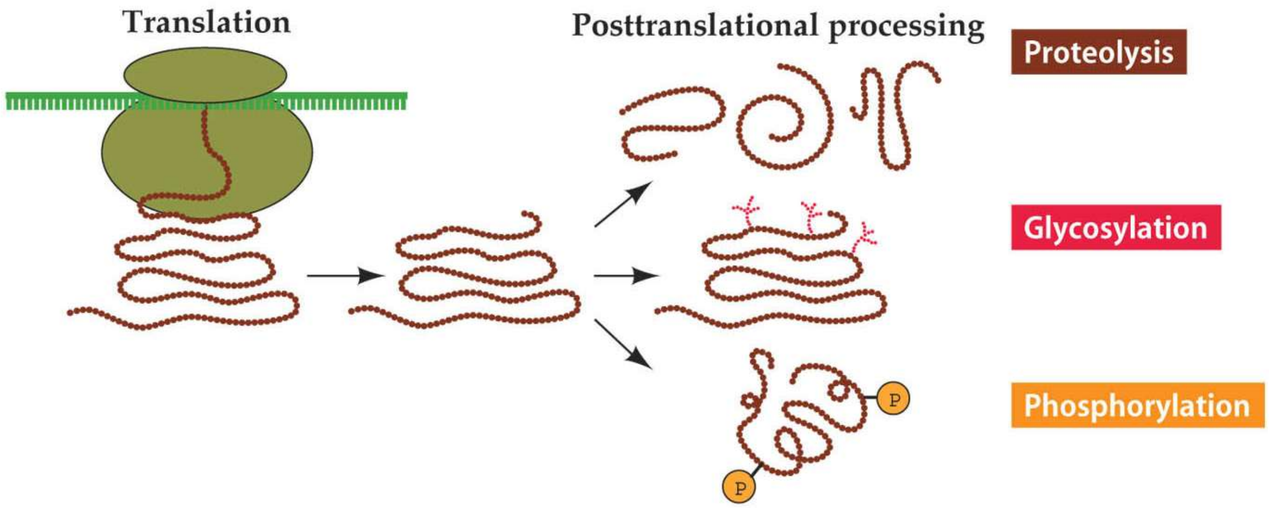


Various proteins belong to different *compartments* of the cell – some even belong *outside* the cell.

Content of UniProt Knowledgebase

- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

Post-translational modifications



Many proteins are *modified* after they have been synthesized in order to become functional.

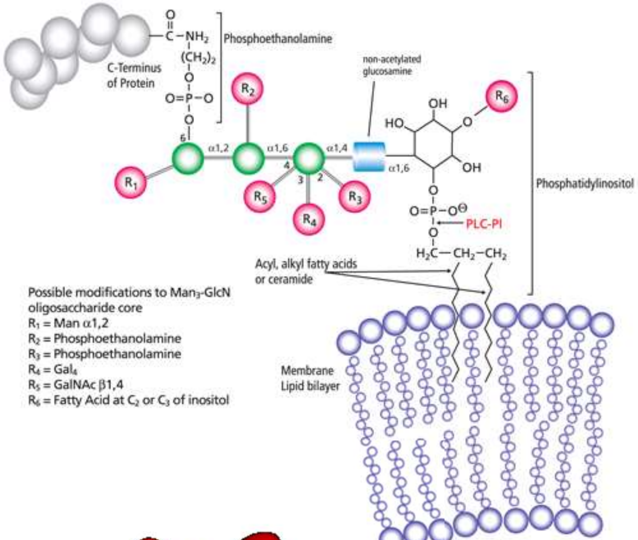
Proteolysis: Cleavage of *signal peptides*, *propeptides* or *initiator methionine.*

Glycosylation: Especially common on the *cell surface*. Plays a role in sorting of proteins to *lysosomes.*

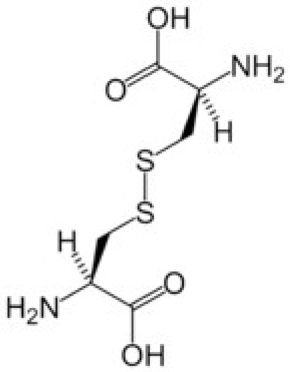
Phosphorylation: Often *reversible*. Regulates the *activity* of many enzymes.

More post-translational modifications

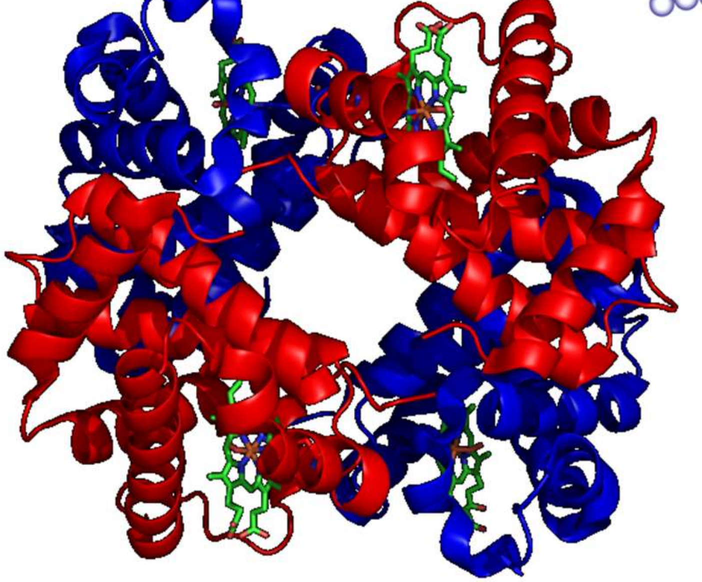
- Lipid anchors
  - (e.g. GPI anchors)



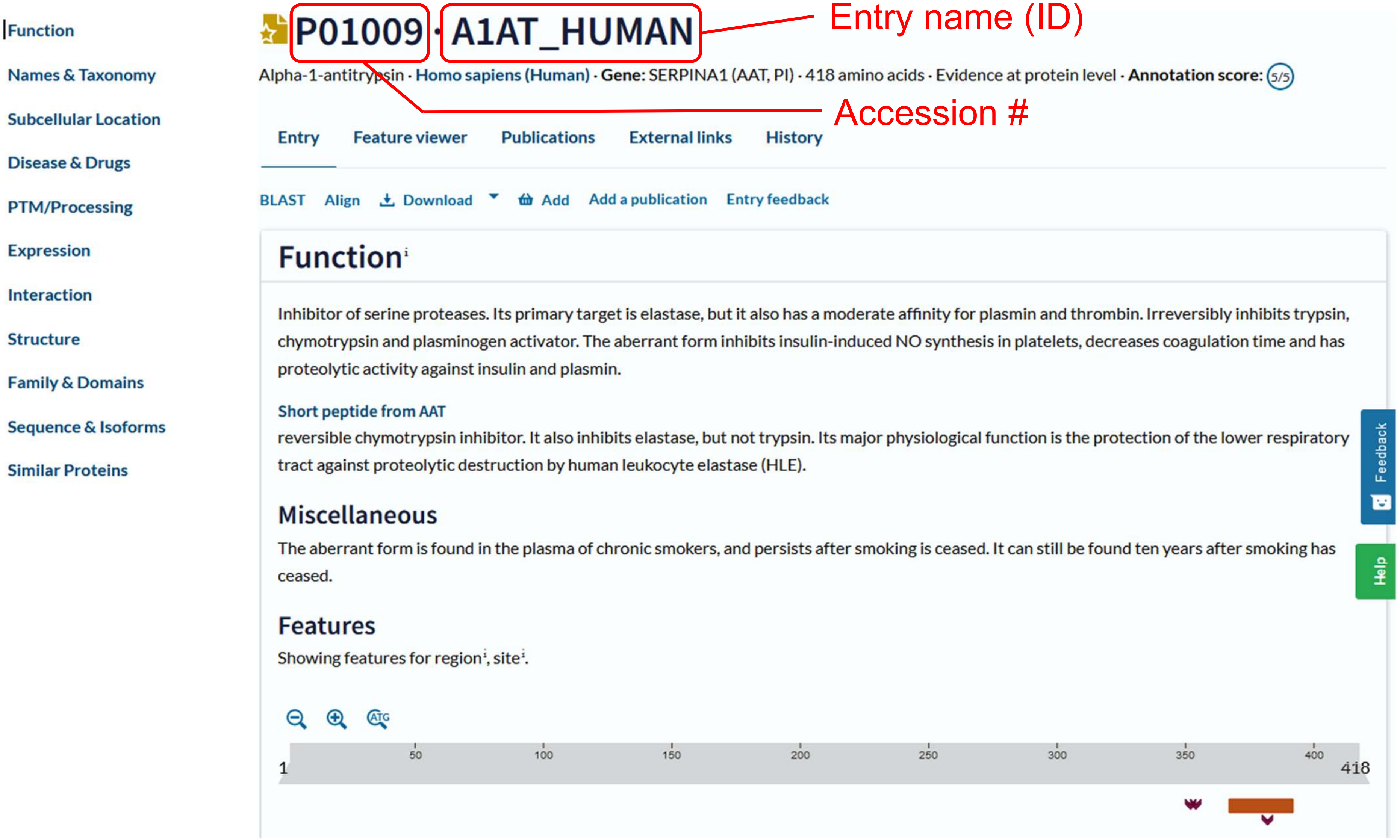
- Disulfide bonds



- Prosthetic groups
  - (*e.g.* metal ions)



UniProt entry, formatted view (new interface)



P01009 · A1AT_HUMAN

Alpha-1-antitrypsin · Homo sapiens (Human) · **Gene:** SERPINA1 (AAT, PI) · 418 amino acids · Evidence at protein level · **Annotation score:** 5/5

Function

Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin.

Short peptide from AAT

reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE).

Miscellaneous

The aberrant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased.

Features

Showing features for region, site.

Entry name (ID)

Accession #

Entry names and accession numbers

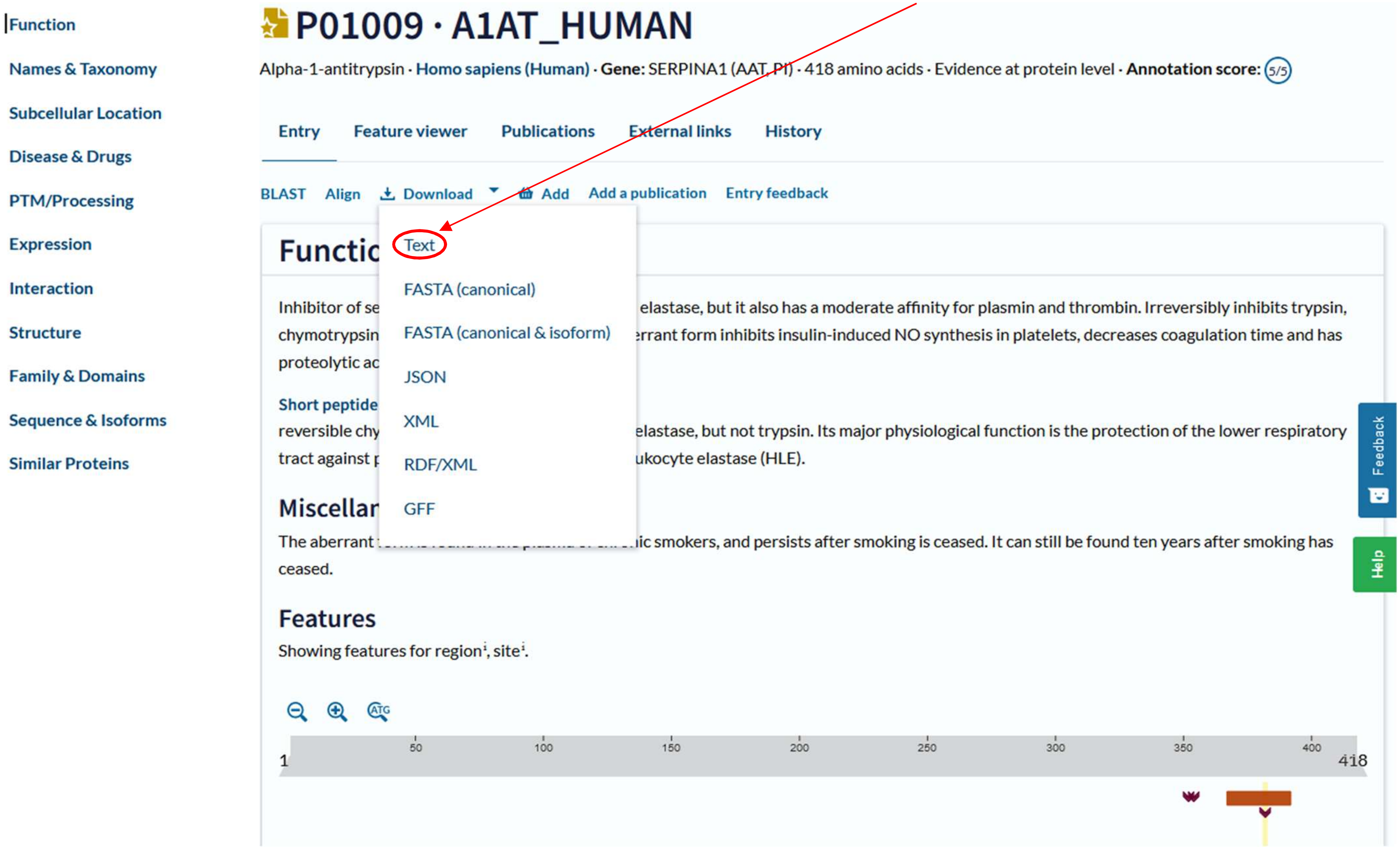
Entry name (UniProt ID / GenBank LOCUS)

Provides a mnemonic identifier for a database entry. One and only one name per entry.

Accession

Provides a *stable* identifier for a database entry (does not change across database versions). One or more accession numbers per entry.

UniProt entry, formatted view



P01009 · A1AT_HUMAN

Alpha-1-antitrypsin · Homo sapiens (Human) · **Gene:** SERPINA1 (AAT, PI) · 418 amino acids · Evidence at protein level · **Annotation score:** 5/5

Functic

Inhibitor of se elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin errant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic ac

Short peptide reversible chy elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against p ukocyte elastase (HLE).

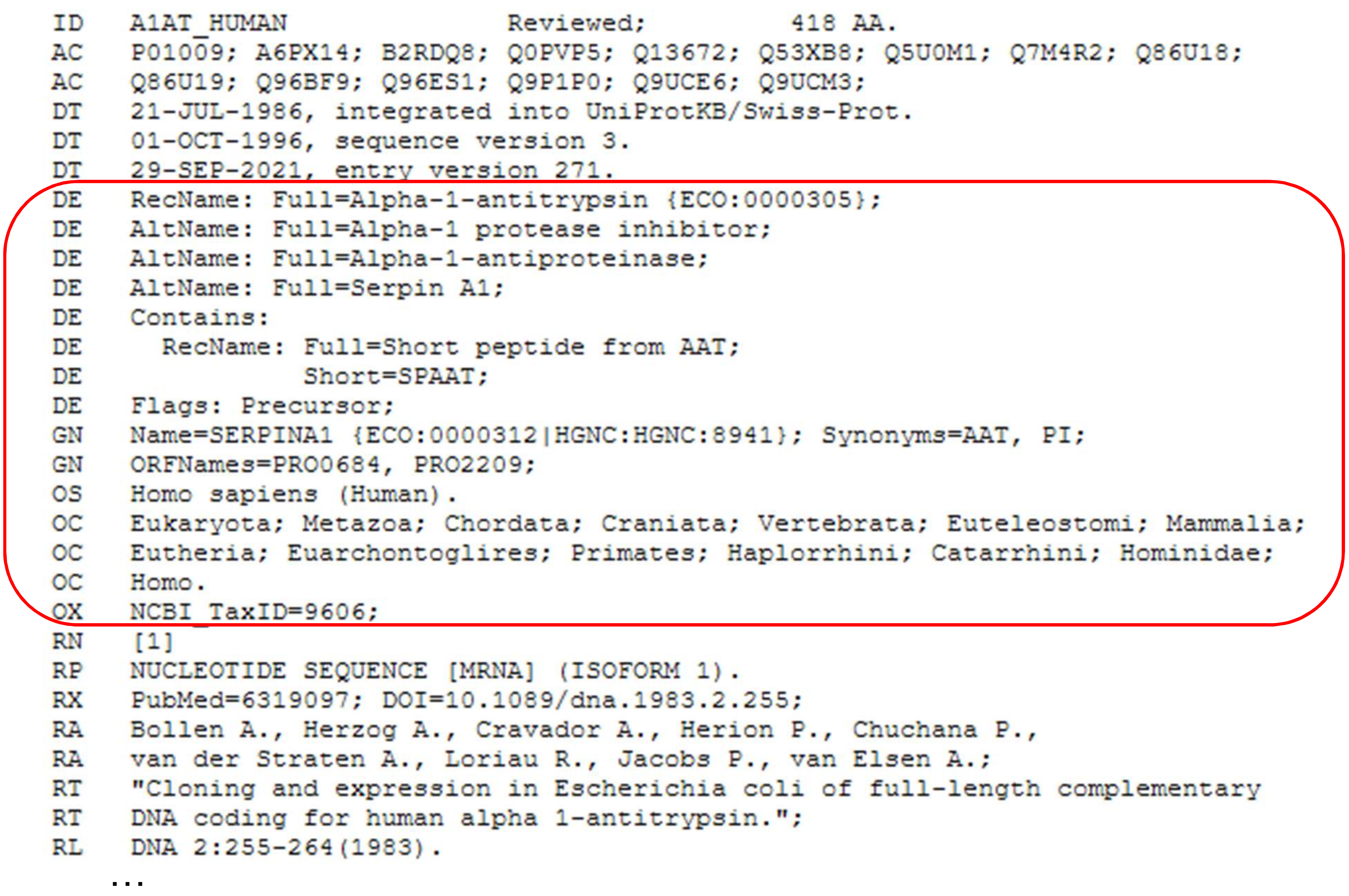
Miscellar

The aberrant ic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased.

Features

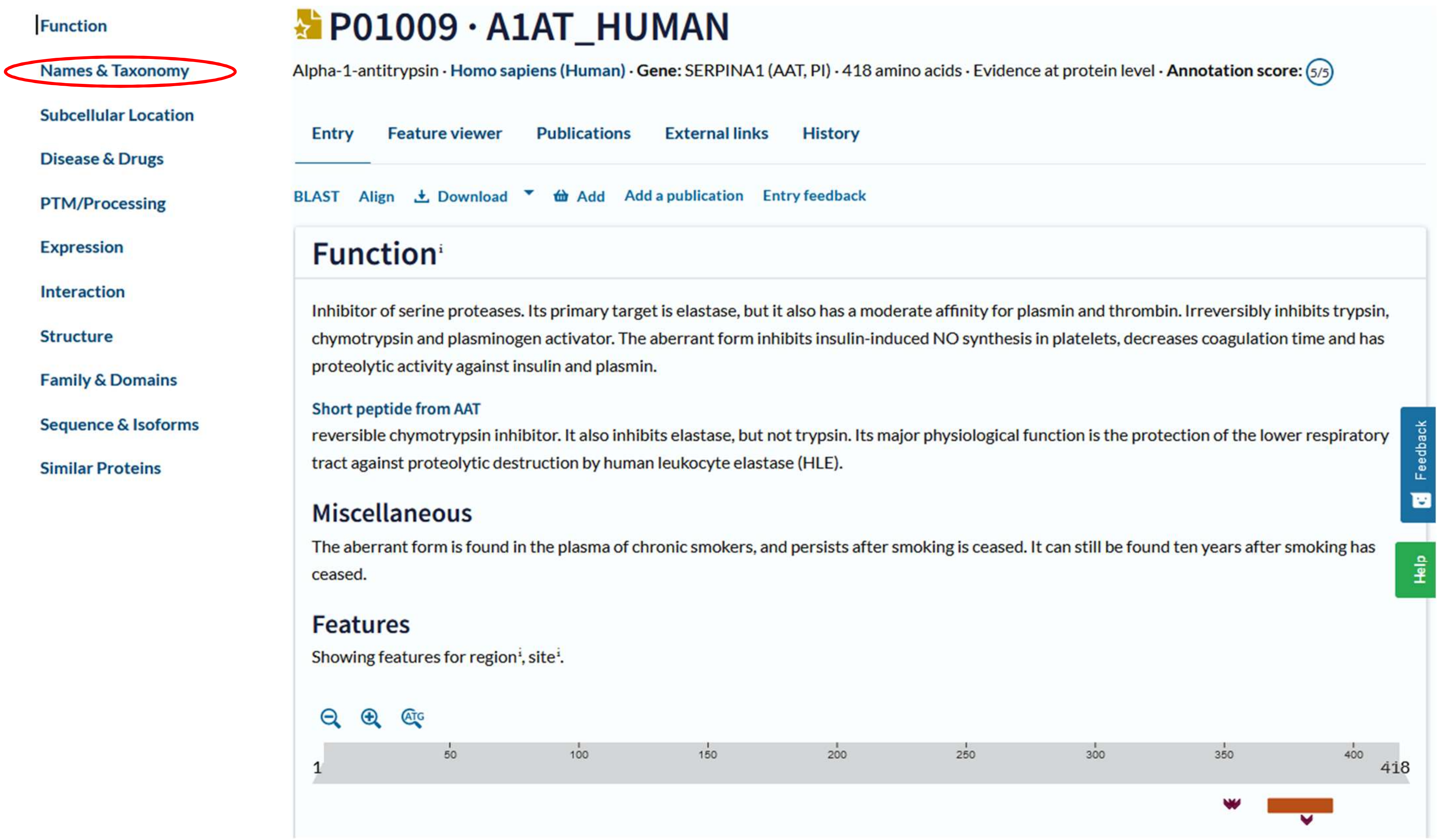
Showing features for region, site.

UniProt entry, text view (flat file)



```
ID   A1AT_HUMAN              Reviewed;         418 AA.
AC   P01009; A6PX14; B2RDQ8; Q0PVP5; Q13672; Q53XB8; Q5U0M1; Q7M4R2; Q86U18;
AC   Q86U19; Q96BF9; Q96ES1; Q9P1P0; Q9UCE6; Q9UCM3;
DT   21-JUL-1986, integrated into UniProtKB/Swiss-Prot.
DT   01-OCT-1996, sequence version 3.
DT   29-SEP-2021, entry version 271.
DE   RecName: Full=Alpha-1-antitrypsin {ECO:0000305};
DE   AltName: Full=Alpha-1 protease inhibitor;
DE   AltName: Full=Alpha-1-antiproteinase;
DE   AltName: Full=Serpin A1;
DE   Contains:
DE     RecName: Full=Short peptide from AAT;
DE              Short=SPAAT;
DE   Flags: Precursor;
GN   Name=SERPINA1 {ECO:0000312|HGNC:HGNC:8941}; Synonyms=AAT, PI;
GN   ORFNames=PRO0684, PRO2209;
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
OX   NCBI_TaxID=9606;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
RX   PubMed=6319097; DOI=10.1089/dna.1983.2.255;
RA   Bollen A., Herzog A., Cravador A., Herion P., Chuchana P.,
RA   van der Straten A., Loriau R., Jacobs P., van Elsen A.;
RT   "Cloning and expression in Escherichia coli of full-length complementary
RT   DNA coding for human alpha 1-antitrypsin.";
RL   DNA 2:255-264(1983).
    ...
```

UniProt entry, formatted view



P01009 · A1AT_HUMAN

Alpha-1-antitrypsin · Homo sapiens (Human) · **Gene:** SERPINA1 (AAT, PI) · 418 amino acids · Evidence at protein level · **Annotation score:** 5/5

Function

Inhibitor of serine proteases. Its primary target is elastase, but it also has a moderate affinity for plasmin and thrombin. Irreversibly inhibits trypsin, chymotrypsin and plasminogen activator. The aberrant form inhibits insulin-induced NO synthesis in platelets, decreases coagulation time and has proteolytic activity against insulin and plasmin.

Short peptide from AAT

reversible chymotrypsin inhibitor. It also inhibits elastase, but not trypsin. Its major physiological function is the protection of the lower respiratory tract against proteolytic destruction by human leukocyte elastase (HLE).

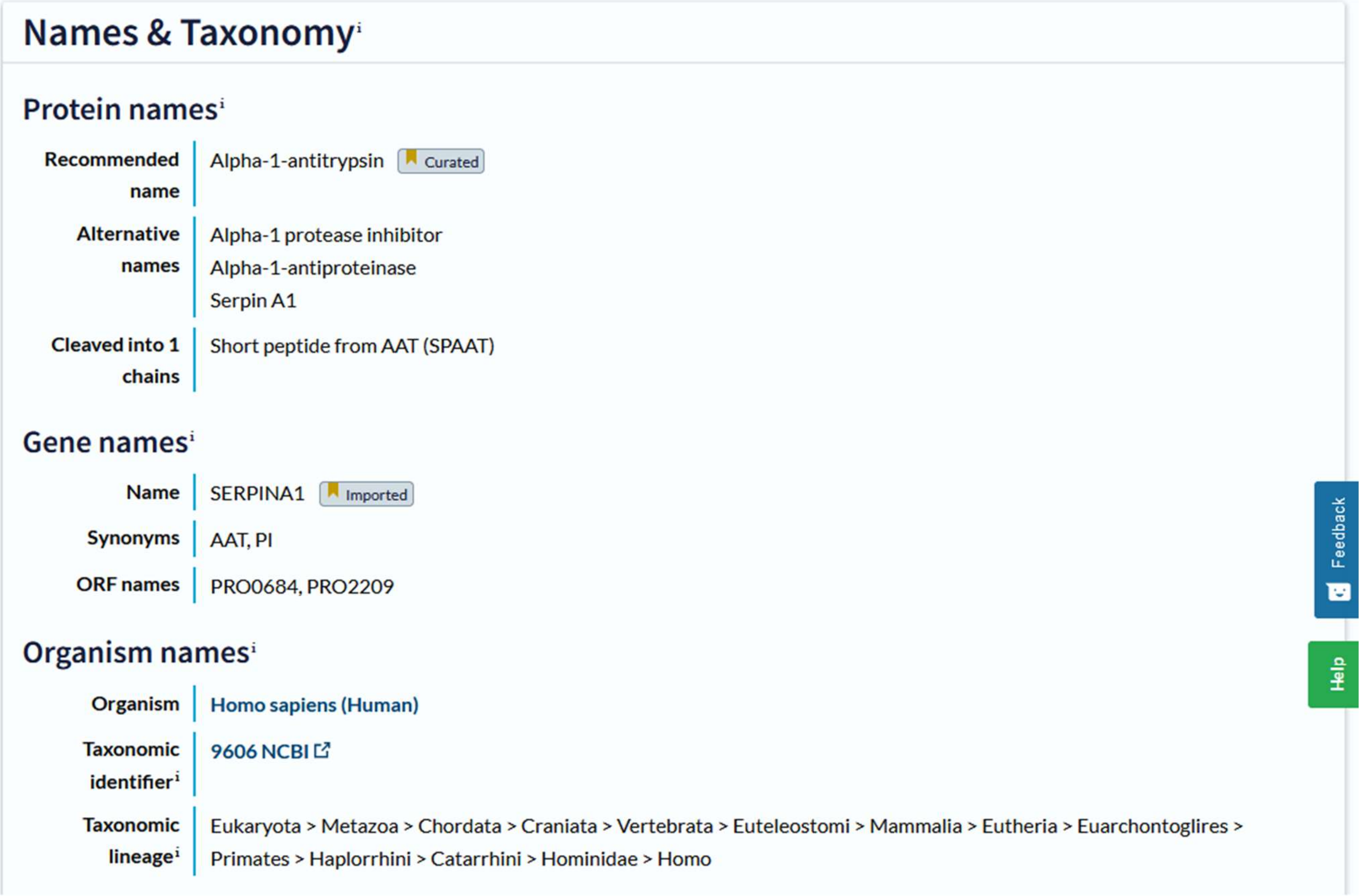
Miscellaneous

The aberrant form is found in the plasma of chronic smokers, and persists after smoking is ceased. It can still be found ten years after smoking has ceased.

Features

Showing features for region, site.

Names & Taxonomy, formatted view



Names & Taxonomy[i]

Protein names[i]

Recommended name	Alpha-1-antitrypsin Curated
Alternative names	Alpha-1 protease inhibitor Alpha-1-antiproteinase Serpin A1
Cleaved into 1 chains	Short peptide from AAT (SPAAT)

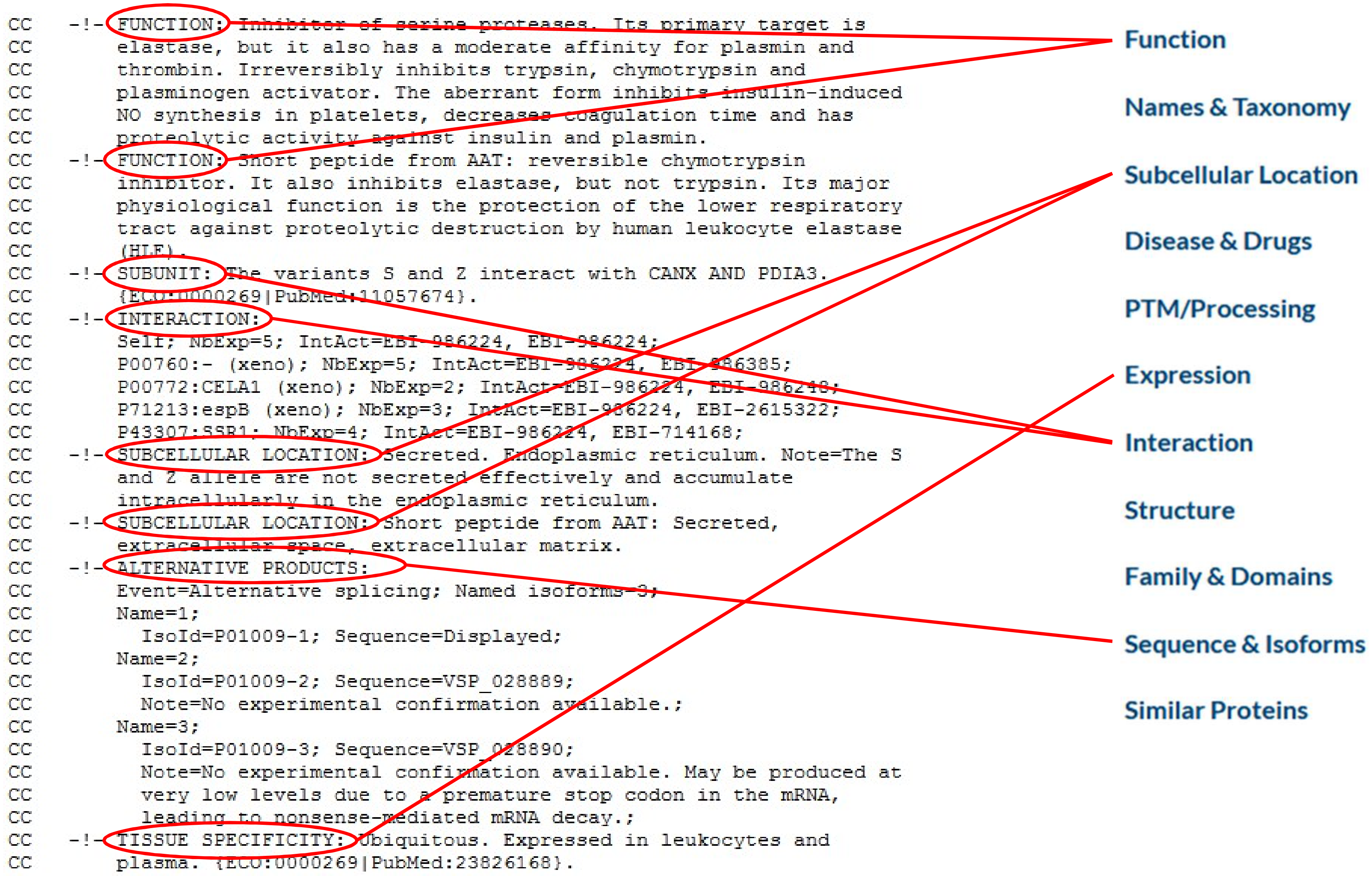
Gene names[i]

Name	SERPINA1 Imported
Synonyms	AAT, PI
ORF names	PRO0684, PRO2209

Organism names[i]

Organism	Homo sapiens (Human)
Taxonomic identifier[i]	9606 NCBI
Taxonomic lineage[i]	Eukaryota > Metazoa > Chordata > Craniata > Vertebrata > Euteleostomi > Mammalia > Eutheria > Euarchontoglires > Primates > Haplorrhini > Catarrhini > Hominidae > Homo

Comments (CC lines)



```
CC   -!- FUNCTION: Inhibitor of serine proteases. Its primary target is
CC       elastase, but it also has a moderate affinity for plasmin and
CC       thrombin. Irreversibly inhibits trypsin, chymotrypsin and
CC       plasminogen activator. The aberrant form inhibits insulin-induced
CC       NO synthesis in platelets, decreases coagulation time and has
CC       proteolytic activity against insulin and plasmin.
CC   -!- FUNCTION: Short peptide from AAT: reversible chymotrypsin
CC       inhibitor. It also inhibits elastase, but not trypsin. Its major
CC       physiological function is the protection of the lower respiratory
CC       tract against proteolytic destruction by human leukocyte elastase
CC       (HLE).
CC   -!- SUBUNIT: The variants S and Z interact with CANX AND PDIA3.
CC       {ECO:0000269|PubMed:11057674}.
CC   -!- INTERACTION:
CC       Self; NbExp=5; IntAct=EBI-986224, EBI-986224;
CC       P00760:- (xeno); NbExp=5; IntAct=EBI-986224, EBI-986385;
CC       P00772:CELA1 (xeno); NbExp=2; IntAct=EBI-986224, EBI-986248;
CC       P71213:espB (xeno); NbExp=3; IntAct=EBI-986224, EBI-2615322;
CC       P43307:SSR1; NbExp=4; IntAct=EBI-986224, EBI-714168;
CC   -!- SUBCELLULAR LOCATION: Secreted. Endoplasmic reticulum. Note=The S
CC       and Z allele are not secreted effectively and accumulate
CC       intracellularly in the endoplasmic reticulum.
CC   -!- SUBCELLULAR LOCATION: Short peptide from AAT: Secreted,
CC       extracellular space, extracellular matrix.
CC   -!- ALTERNATIVE PRODUCTS:
CC       Event=Alternative splicing; Named isoforms=3;
CC       Name=1;
CC         IsoId=P01009-1; Sequence=Displayed;
CC       Name=2;
CC         IsoId=P01009-2; Sequence=VSP_028889;
CC         Note=No experimental confirmation available.;
CC       Name=3;
CC         IsoId=P01009-3; Sequence=VSP_028890;
CC         Note=No experimental confirmation available. May be produced at
CC         very low levels due to a premature stop codon in the mRNA,
CC         leading to nonsense-mediated mRNA decay.;
CC   -!- TISSUE SPECIFICITY: Ubiquitous. Expressed in leukocytes and
CC       plasma. {ECO:0000269|PubMed:23826168}.
```

Function

Names & Taxonomy

Subcellular Location

Disease & Drugs

PTM/Processing

Expression

Interaction

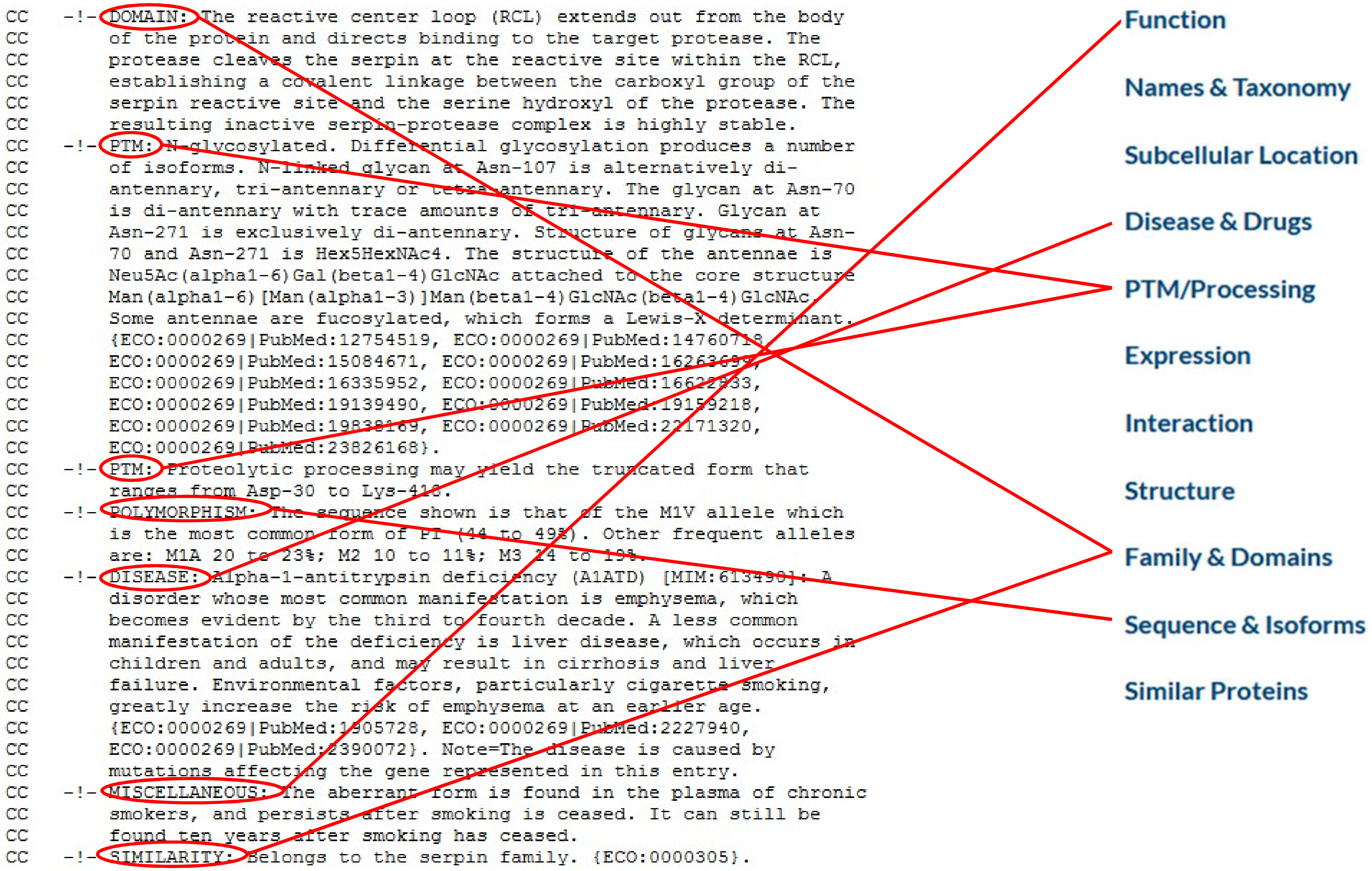
Structure

Family & Domains

Sequence & Isoforms

Similar Proteins

Comments (CC lines), continued



```
CC   -!- DOMAIN: The reactive center loop (RCL) extends out from the body
CC       of the protein and directs binding to the target protease. The
CC       protease cleaves the serpin at the reactive site within the RCL,
CC       establishing a covalent linkage between the carboxyl group of the
CC       serpin reactive site and the serine hydroxyl of the protease. The
CC       resulting inactive serpin-protease complex is highly stable.
CC   -!- PTM: N-glycosylated. Differential glycosylation produces a number
CC       of isoforms. N-linked glycan at Asn-107 is alternatively di-
CC       antennary, tri-antennary or tetra-antennary. The glycan at Asn-70
CC       is di-antennary with trace amounts of tri-antennary. Glycan at
CC       Asn-271 is exclusively di-antennary. Structure of glycans at Asn-
CC       70 and Asn-271 is Hex5HexNAc4. The structure of the antennae is
CC       Neu5Ac(alpha1-6)Gal(beta1-4)GlcNAc attached to the core structure
CC       Man(alpha1-6)[Man(alpha1-3)]Man(beta1-4)GlcNAc(beta1-4)GlcNAc.
CC       Some antennae are fucosylated, which forms a Lewis-X determinant.
CC       {ECO:0000269|PubMed:12754519, ECO:0000269|PubMed:14760718,
CC       ECO:0000269|PubMed:15084671, ECO:0000269|PubMed:16263699,
CC       ECO:0000269|PubMed:16335952, ECO:0000269|PubMed:16622933,
CC       ECO:0000269|PubMed:19139490, ECO:0000269|PubMed:19159218,
CC       ECO:0000269|PubMed:19838169, ECO:0000269|PubMed:22171320,
CC       ECO:0000269|PubMed:23826168}.
CC   -!- PTM: Proteolytic processing may yield the truncated form that
CC       ranges from Asp-30 to Lys-418.
CC   -!- POLYMORPHISM: The sequence shown is that of the M1V allele which
CC       is the most common form of PI (44 to 49%). Other frequent alleles
CC       are: M1A 20 to 23%; M2 10 to 11%; M3 14 to 19%.
CC   -!- DISEASE: Alpha-1-antitrypsin deficiency (A1ATD) [MIM:613490]: A
CC       disorder whose most common manifestation is emphysema, which
CC       becomes evident by the third to fourth decade. A less common
CC       manifestation of the deficiency is liver disease, which occurs in
CC       children and adults, and may result in cirrhosis and liver
CC       failure. Environmental factors, particularly cigarette smoking,
CC       greatly increase the risk of emphysema at an earlier age.
CC       {ECO:0000269|PubMed:1905728, ECO:0000269|PubMed:2227940,
CC       ECO:0000269|PubMed:2390072}. Note=The disease is caused by
CC       mutations affecting the gene represented in this entry.
CC   -!- MISCELLANEOUS: The aberrant form is found in the plasma of chronic
CC       smokers, and persists after smoking is ceased. It can still be
CC       found ten years after smoking has ceased.
CC   -!- SIMILARITY: Belongs to the serpin family. {ECO:0000305}.
```

Function

Names & Taxonomy

Subcellular Location

Disease & Drugs

PTM/Processing

Expression

Interaction

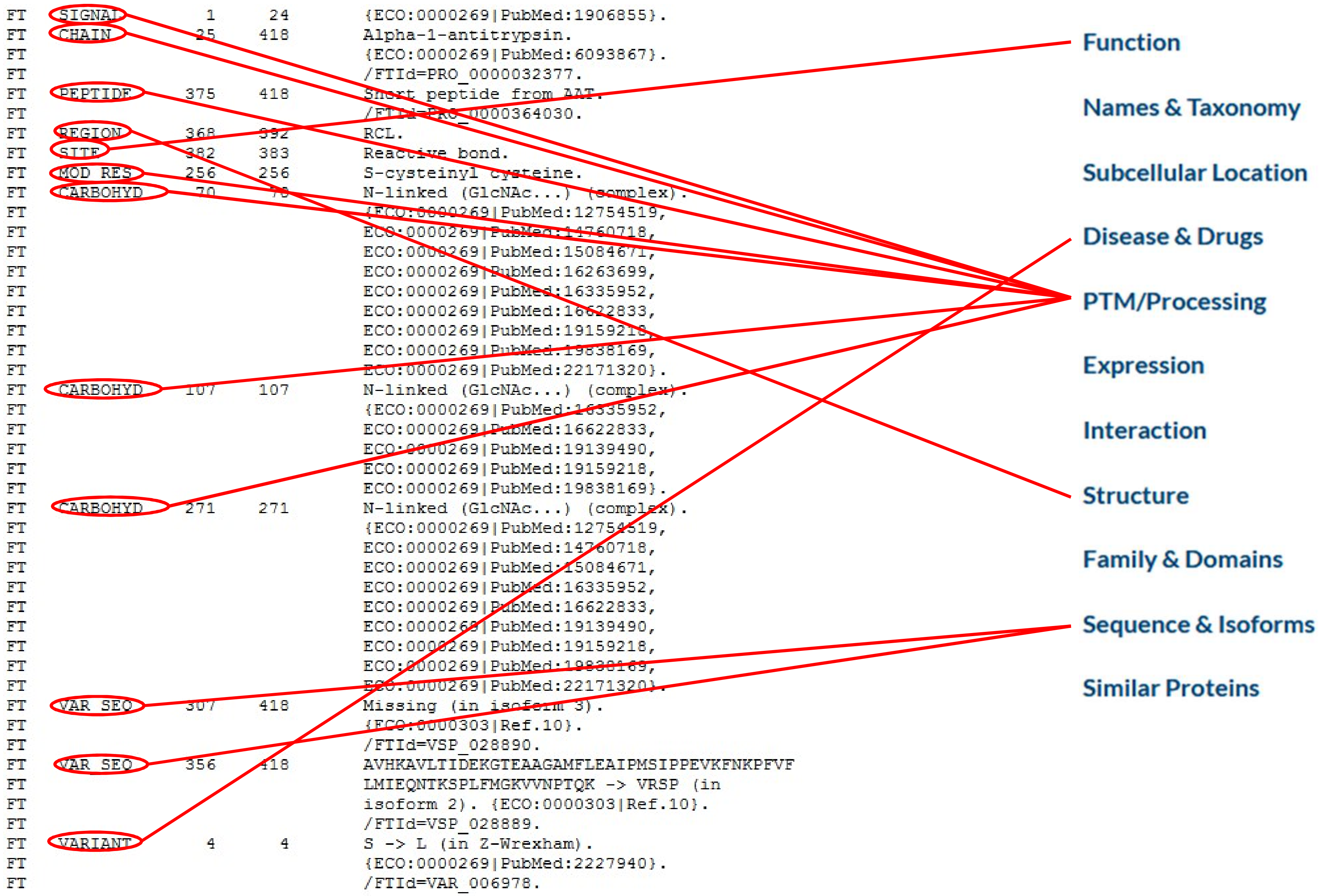
Structure

Family & Domains

Sequence & Isoforms

Similar Proteins

Feature table (FT lines)



```
FT   SIGNAL        1     24       {ECO:0000269|PubMed:1906855}.
FT   CHAIN        25    418       Alpha-1-antitrypsin.
FT                                {ECO:0000269|PubMed:6093867}.
FT                                /FTId=PRO_0000032377.
FT   PEPTIDE     375    418       Short peptide from AAT.
FT                                /FTId=PRO_0000364030.
FT   REGION      368    392       RCL.
FT   SITE        382    383       Reactive bond.
FT   MOD_RES     256    256       S-cysteinyl cysteine.
FT   CARBOHYD     70     70       N-linked (GlcNAc...) (complex).
FT                                {ECO:0000269|PubMed:12754519,
FT                                ECO:0000269|PubMed:14760718,
FT                                ECO:0000269|PubMed:15084671,
FT                                ECO:0000269|PubMed:16263699,
FT                                ECO:0000269|PubMed:16335952,
FT                                ECO:0000269|PubMed:16622833,
FT                                ECO:0000269|PubMed:19159218,
FT                                ECO:0000269|PubMed:19838169,
FT                                ECO:0000269|PubMed:22171320}.
FT   CARBOHYD    107    107       N-linked (GlcNAc...) (complex).
FT                                {ECO:0000269|PubMed:16335952,
FT                                ECO:0000269|PubMed:16622833,
FT                                ECO:0000269|PubMed:19139490,
FT                                ECO:0000269|PubMed:19159218,
FT                                ECO:0000269|PubMed:19838169}.
FT   CARBOHYD    271    271       N-linked (GlcNAc...) (complex).
FT                                {ECO:0000269|PubMed:12754519,
FT                                ECO:0000269|PubMed:14760718,
FT                                ECO:0000269|PubMed:15084671,
FT                                ECO:0000269|PubMed:16335952,
FT                                ECO:0000269|PubMed:16622833,
FT                                ECO:0000269|PubMed:19139490,
FT                                ECO:0000269|PubMed:19159218,
FT                                ECO:0000269|PubMed:19838169,
FT                                ECO:0000269|PubMed:22171320}.
FT   VAR_SEQ     307    418       Missing (in isoform 3).
FT                                {ECO:0000303|Ref.10}.
FT                                /FTId=VSP_028890.
FT   VAR_SEQ     356    418       AVHKAVLTIDEKGTEAAGAMFLEAIPMSIPPEVKFNKPFVF
FT                                LMIEQNTKSPLFMGKVVNPTQK -> VRSP (in
FT                                isoform 2). {ECO:0000303|Ref.10}.
FT                                /FTId=VSP_028889.
FT   VARIANT       4      4       S -> L (in Z-Wrexham).
FT                                {ECO:0000269|PubMed:2227940}.
FT                                /FTId=VAR_006978.
```

Function

Names & Taxonomy

Subcellular Location

Disease & Drugs

PTM/Processing

Expression

Interaction

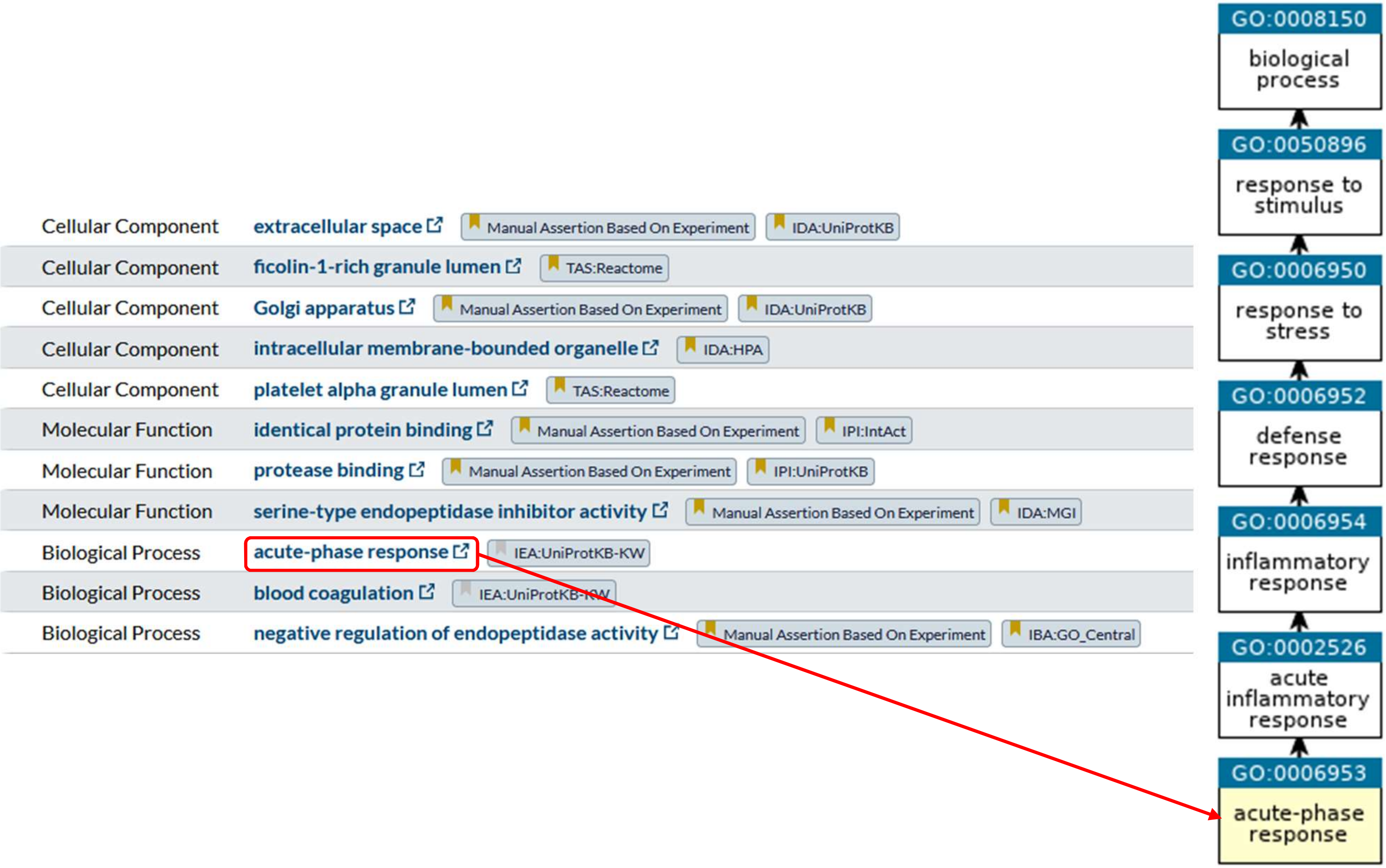
Structure

Family & Domains

Sequence & Isoforms

Similar Proteins

Gene Ontology (GO)



Cellular Component	extracellular space	Manual Assertion Based On Experiment	IDA:UniProtKB
Cellular Component	ficolin-1-rich granule lumen	TAS:Reactome	
Cellular Component	Golgi apparatus	Manual Assertion Based On Experiment	IDA:UniProtKB
Cellular Component	intracellular membrane-bounded organelle	IDA:HPA	
Cellular Component	platelet alpha granule lumen	TAS:Reactome	
Molecular Function	identical protein binding	Manual Assertion Based On Experiment	IPI:IntAct
Molecular Function	protease binding	Manual Assertion Based On Experiment	IPI:UniProtKB
Molecular Function	serine-type endopeptidase inhibitor activity	Manual Assertion Based On Experiment	IDA:MGI
Biological Process	acute-phase response	IEA:UniProtKB-KW	
Biological Process	blood coagulation	IEA:UniProtKB-KW	
Biological Process	negative regulation of endopeptidase activity	Manual Assertion Based On Experiment	IBA:GO_Central

GO:0008150 biological process

GO:0050896 response to stimulus

GO:0006950 response to stress

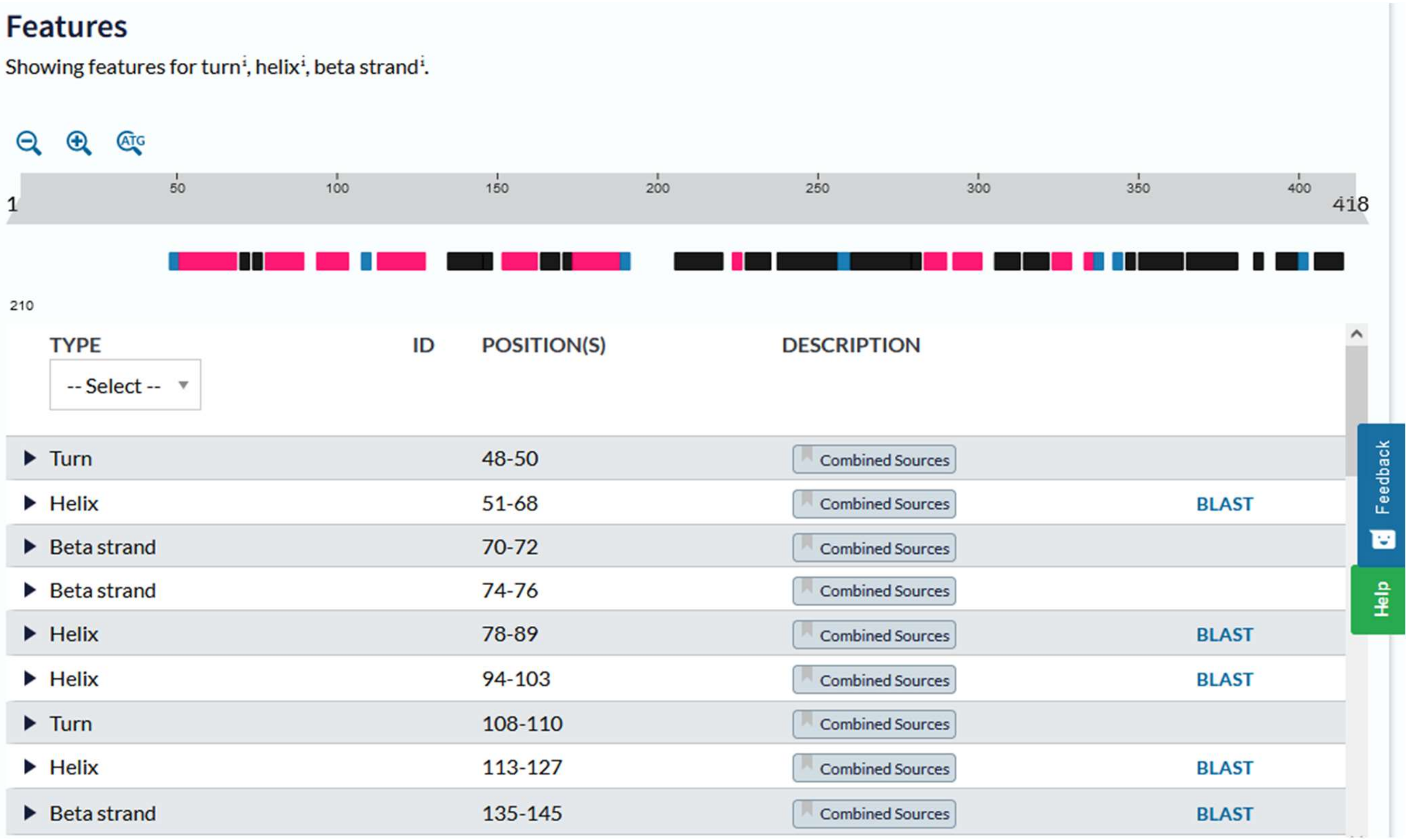
GO:0006952 defense response

GO:0006954 inflammatory response

GO:0002526 acute inflammatory response

GO:0006953 acute-phase response

Secondary structure (Feature Table)



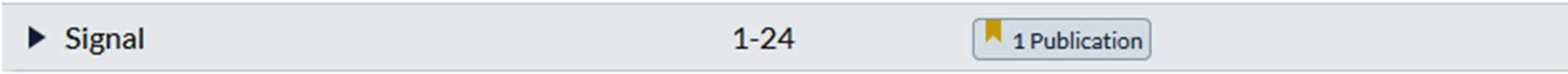
Features

Showing features for turn, helix, beta strand.

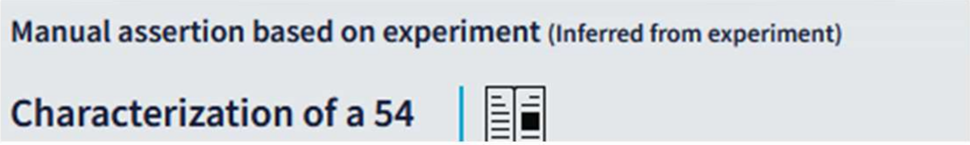
TYPE -- Select --	ID	POSITION(S)	DESCRIPTION	
▶ Turn		48-50	Combined Sources	
▶ Helix		51-68	Combined Sources	BLAST
▶ Beta strand		70-72	Combined Sources	
▶ Beta strand		74-76	Combined Sources	
▶ Helix		78-89	Combined Sources	BLAST
▶ Helix		94-103	Combined Sources	BLAST
▶ Turn		108-110	Combined Sources	
▶ Helix		113-127	Combined Sources	BLAST
▶ Beta strand		135-145	Combined Sources	BLAST

Evidence (Comments, Feature Table)

Experimental (A1AT_HUMAN):



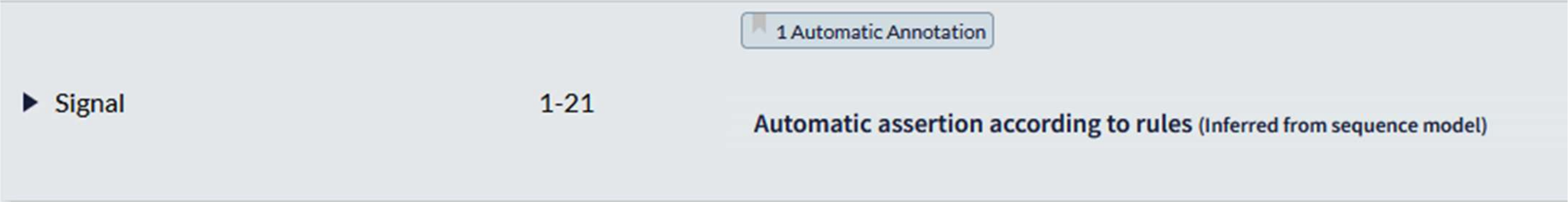
▶ Signal 1-24 1 Publication



Manual assertion based on experiment (Inferred from experiment)

Characterization of a 54

Predicted (VWC2L_HUMAN):



▶ Signal 1-21 1 Automatic Annotation

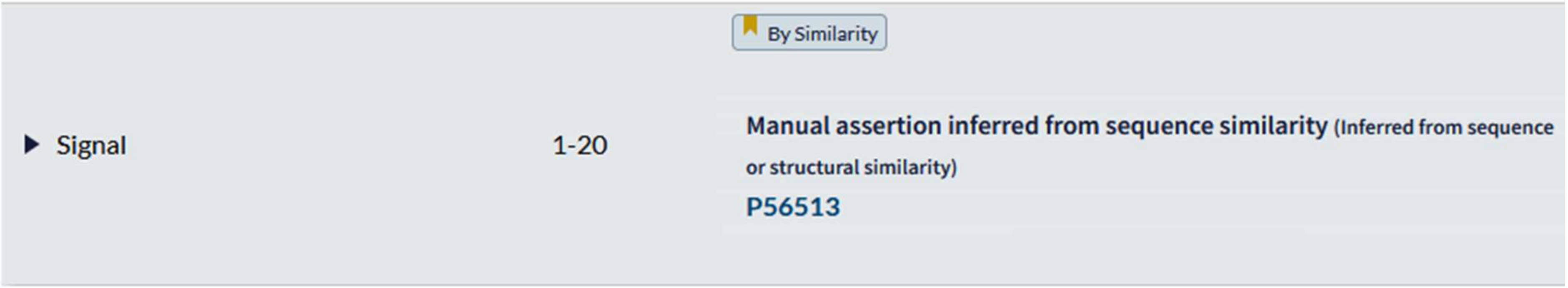
Automatic assertion according to rules (Inferred from sequence model)

patient.

Tanaka N., Sekiya S.,

Takamizawa H., Kato N.,

By similarity (PLM_HUMAN):

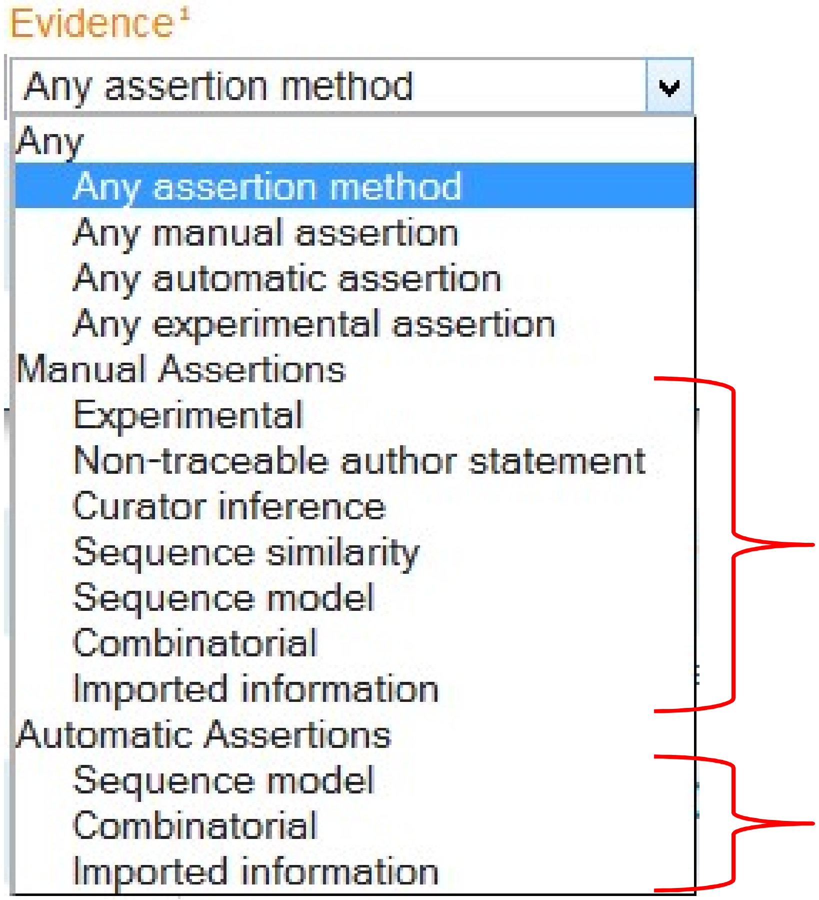


▶ Signal 1-20 By Similarity

Manual assertion inferred from sequence similarity (Inferred from sequence or structural similarity)

P56513

Evidence types in UniProt



Evidence¹

Any assertion method

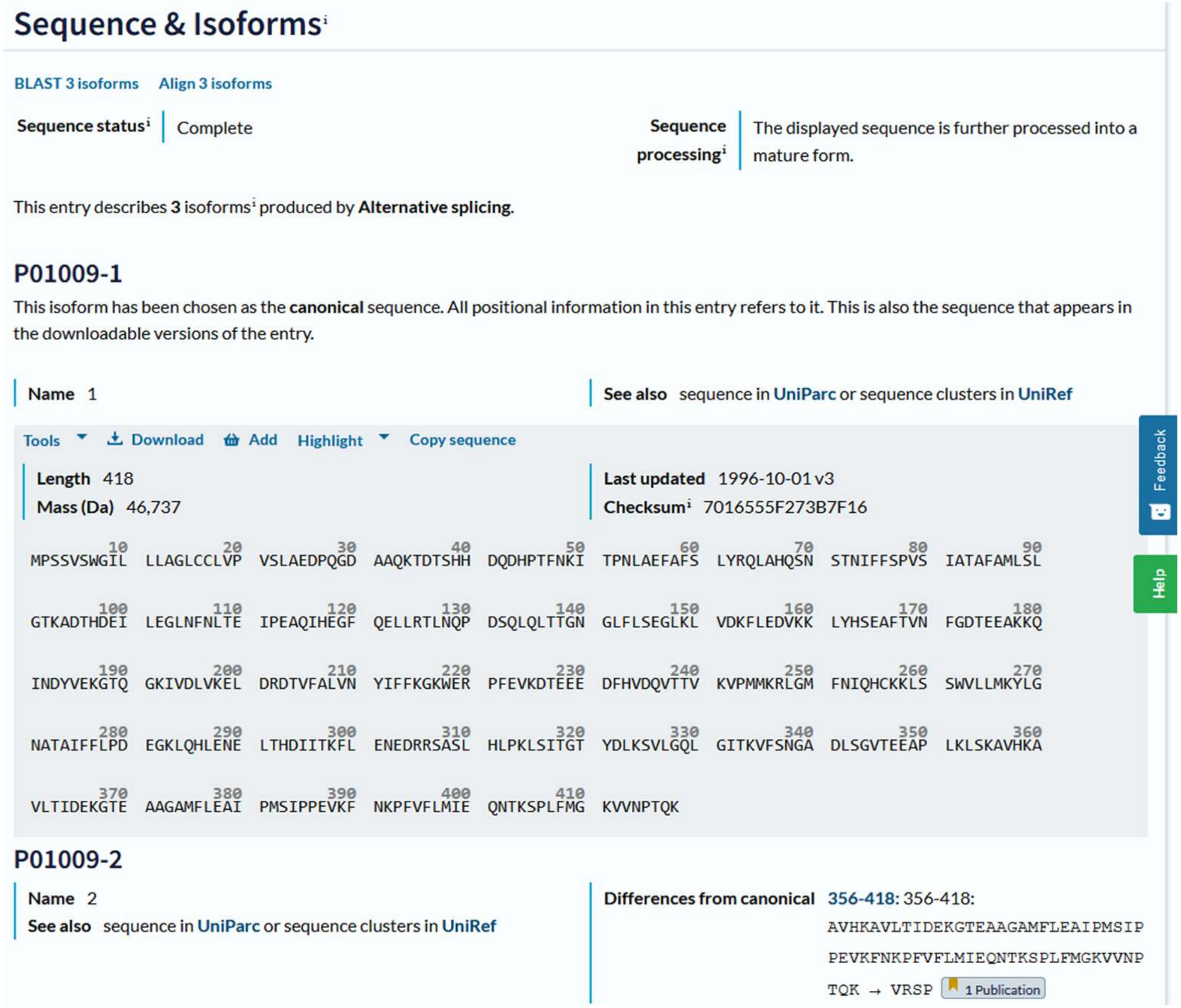
- Any
  - Any assertion method
  - Any manual assertion
  - Any automatic assertion
  - Any experimental assertion
- Manual Assertions
  - Experimental
  - Non-traceable author statement
  - Curator inference
  - Sequence similarity
  - Sequence model
  - Combinatorial
  - Imported information
- Automatic Assertions
  - Sequence model
  - Combinatorial
  - Imported information

Used in Swiss-Prot

Used in TrEMBL

See also http://www.uniprot.org/help/evidences

UniProt entry, sequence(s)



Sequence & Isoforms[i]

BLAST 3 isoforms Align 3 isoforms

Sequence status[i] Complete

Sequence processing[i] The displayed sequence is further processed into a mature form.

This entry describes **3** isoforms[i] produced by **Alternative splicing**.

P01009-1

This isoform has been chosen as the **canonical** sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Name 1

See also sequence in UniParc or sequence clusters in UniRef

Length 418
Mass (Da) 46,737

Last updated 1996-10-01 v3
Checksum[i] 7016555F273B7F16

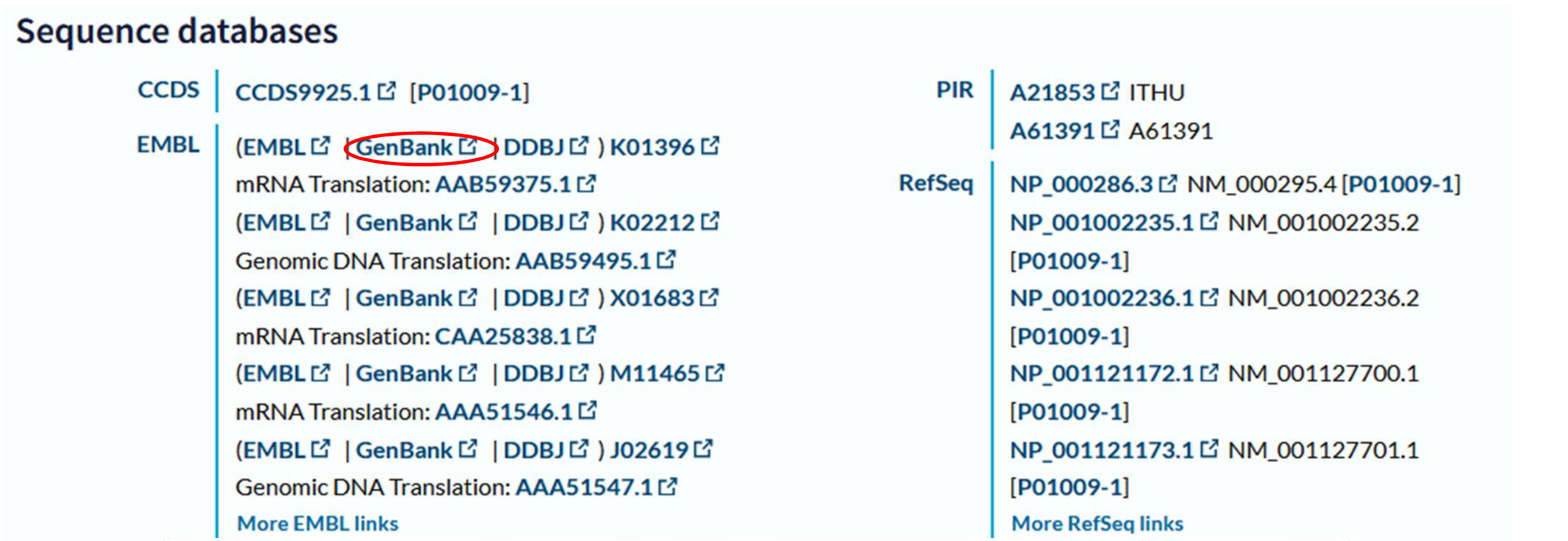
```
        10         20         30         40         50         60         70         80         90
MPSSVSWGIL LLAGLCCLVP VSLAEDPQGD AAQKTDTSHH DQDHPTFNKI TPNLAEFAFS LYRQLAHQSN STNIFFSPVS IATAFAMLSL

       100        110        120        130        140        150        160        170        180
GTKADTHDEI LEGLNFNLTE IPEAQIHEGF QELLRTLNQP DSQLQLTTGN GLFLSEGLKL VDKFLEDVKK LYHSEAFTVN FGDTEEAKKQ

       190        200        210        220        230        240        250        260        270
INDYVEKGTQ GKIVDLVKEL DRDTVFALVN YIFFKGKWER PFEVKDTEEE DFHVDQVTTV KVPMMKRLGM FNIQHCKKLS SWVLLMKYLG

       280        290        300        310        320        330        340        350        360
NATAIFFLPD EGKLQHLENE LTHDIITKFL ENEDRRSASL HLPKLSITGT YDLKSVLGQL GITKVFSNGA DLSGVTEEAP LKLSKAVHKA

       370        380        390        400        410
VLTIDEKGTE AAGAMFLEAI PMSIPPEVKF NKPFVFLMIE QNTKSPLFMG KVVNPTQK
```

P01009-2

Name 2
See also sequence in UniParc or sequence clusters in UniRef

Differences from canonical 356-418: 356-418:
AVHKAVLTIDEKGTEAAGAMFLEAIPMSIP
PEVKFNKPFVFLMIEQNTKSPLFMGKVVNP
TQK → VRSP 1 Publication

Cross-references, nucleotide sequences



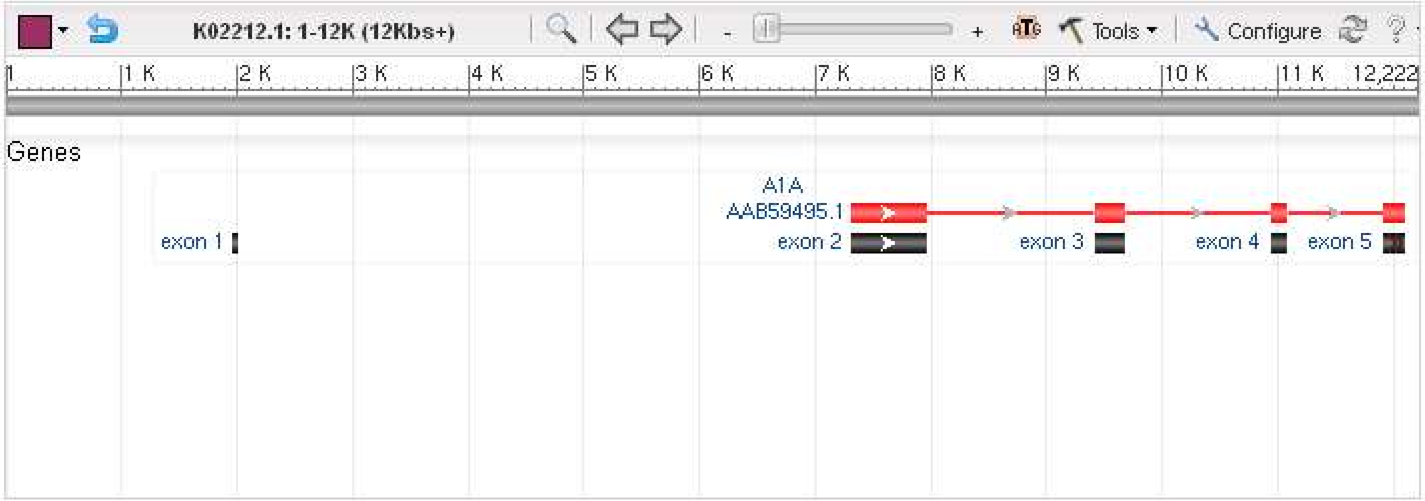
Sequence databases

CCDS
CCDS9925.1 [P01009-1]

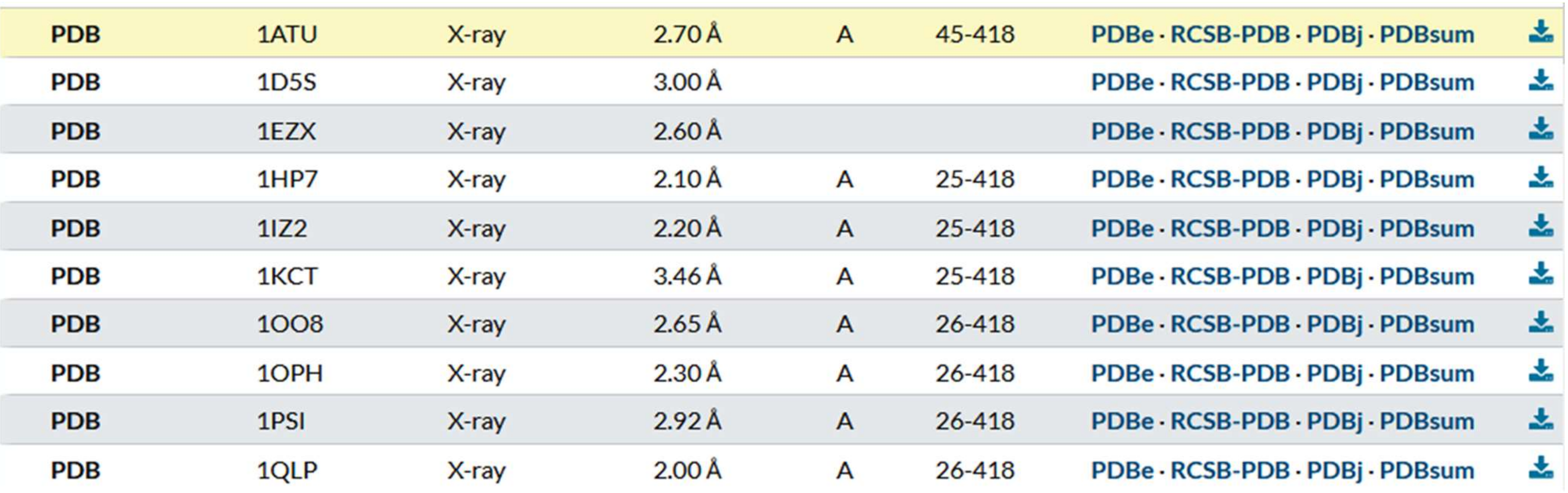
EMBL
(EMBL | GenBank | DDBJ) K01396
mRNA Translation: AAB59375.1
(EMBL | GenBank | DDBJ) K02212
Genomic DNA Translation: AAB59495.1
(EMBL | GenBank | DDBJ) X01683
mRNA Translation: CAA25838.1
(EMBL | GenBank | DDBJ) M11465
mRNA Translation: AAA51546.1
(EMBL | GenBank | DDBJ) J02619
Genomic DNA Translation: AAA51547.1
More EMBL links

PIR
A21853 ITHU
A61391 A61391

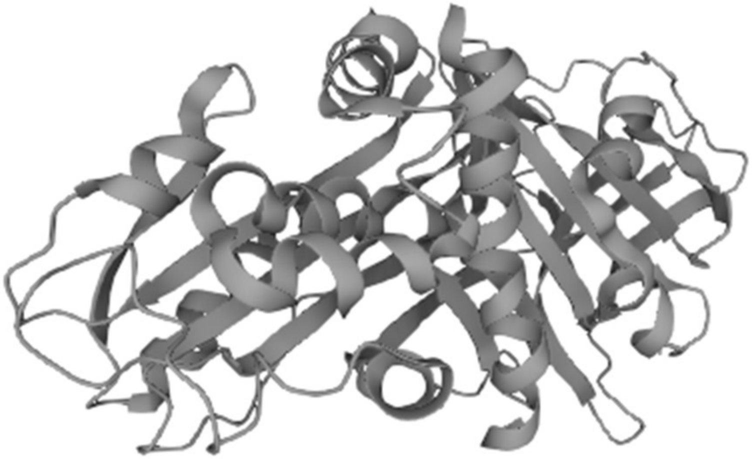
RefSeq
NP_000286.3 NM_000295.4 [P01009-1]
NP_001002235.1 NM_001002235.2 [P01009-1]
NP_001002236.1 NM_001002236.2 [P01009-1]
NP_001121172.1 NM_001127700.1 [P01009-1]
NP_001121173.1 NM_001127701.1 [P01009-1]
More RefSeq links



Cross-references, 3D structure



PDB	1ATU	X-ray	2.70 Å	A	45-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1D5S	X-ray	3.00 Å			PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1EZX	X-ray	2.60 Å			PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1HP7	X-ray	2.10 Å	A	25-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1IZ2	X-ray	2.20 Å	A	25-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1KCT	X-ray	3.46 Å	A	25-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1OO8	X-ray	2.65 Å	A	26-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1OPH	X-ray	2.30 Å	A	26-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1PSI	X-ray	2.92 Å	A	26-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇
PDB	1QLP	X-ray	2.00 Å	A	26-418	PDBe · RCSB-PDB · PDBj · PDBsum	⬇



Cross-references

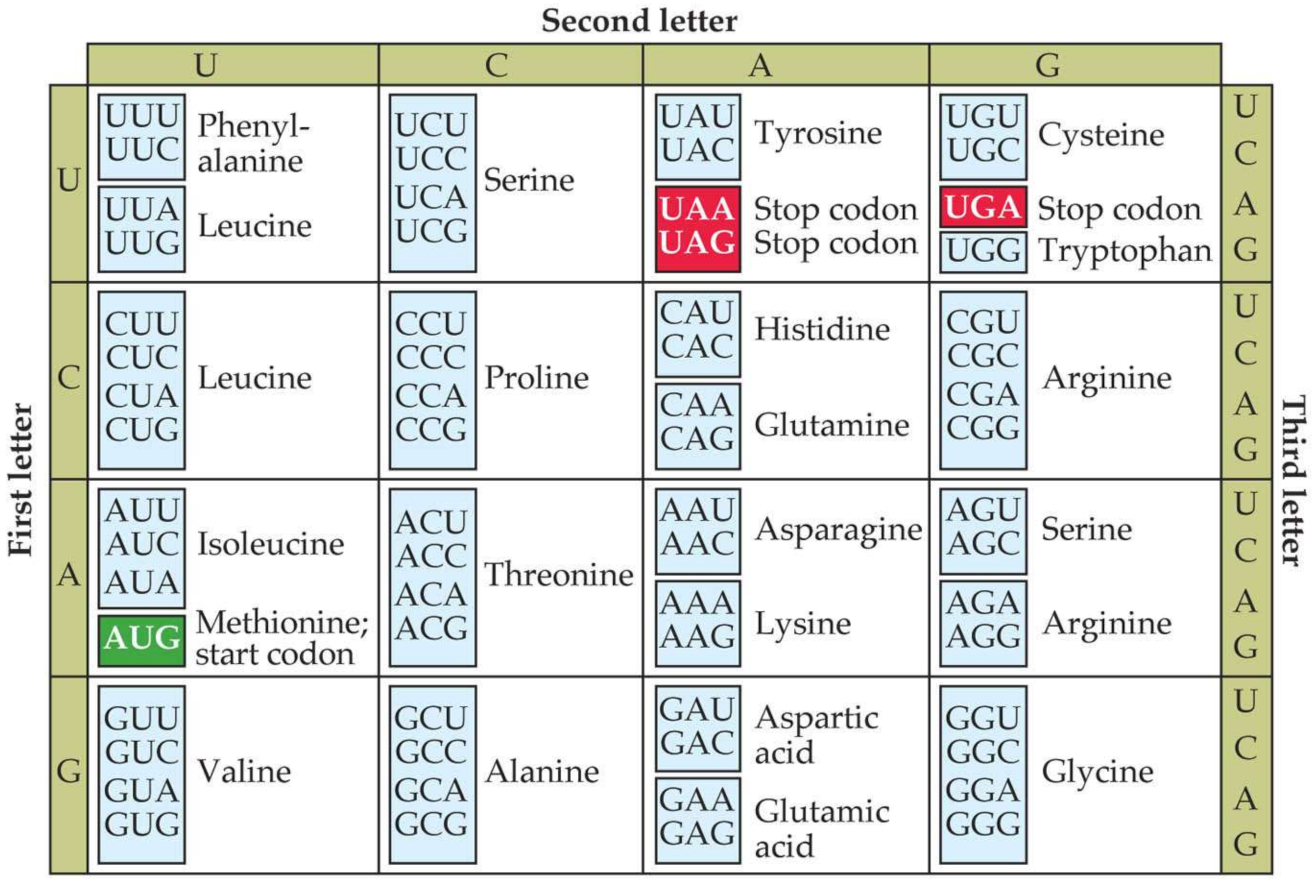
Other databases linked from UniProt

(there are ~100 in total):

- Nucleotide sequences
- 3D structure
- Protein-protein interactions
- Enzymatic activities and pathways
- Gene expression (microarrays and 2D-PAGE)
- Ontologies
- Families and domains
- Organism specific databases

Translation
and
Reading Frames

The genetic code



	Second letter				
First letter	U	C	A	G	Third letter
U	UUU Phenylalanine	UCU Serine	UAU Tyrosine	UGU Cysteine	U
	UUC	UCC	UAC	UGC	C
	UUA Leucine	UCA	UAA Stop codon	UGA Stop codon	A
	UUG	UCG	UAG Stop codon	UGG Tryptophan	G
C	CUU Leucine	CCU Proline	CAU Histidine	CGU Arginine	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Glutamine	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Isoleucine	ACU Threonine	AAU Asparagine	AGU Serine	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lysine	AGA Arginine	A
	AUG Methionine; start codon	ACG	AAG	AGG	G
G	GUU Valine	GCU Alanine	GAU Aspartic acid	GGU Glycine	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glutamic acid	GGA	A
	GUG	GCG	GAG	GGG	G

- Degenerate (*redundant*) but not ambiguous
- *Almost* universal (deviations found in mitochondria)

Reading Frames 1

A piece of an mRNA-strand:

```
5'                                                                3'
augcccaagcugaauagcguagagggguuuucaucauuugaggacgauguauaa
```

can be divided into triplets (*codons*) in three ways:

```
1 aug ccc aag cug aau agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  ugc cca agc uga aua gcg uag agg ggu uuu cau cau uug agg acg aug uau
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gcu gaa uag cgu aga ggg guu uuc auc auu uga gga cga ugu aua
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Each possible set of triplets is called a *reading frame.*

Reading Frames 2

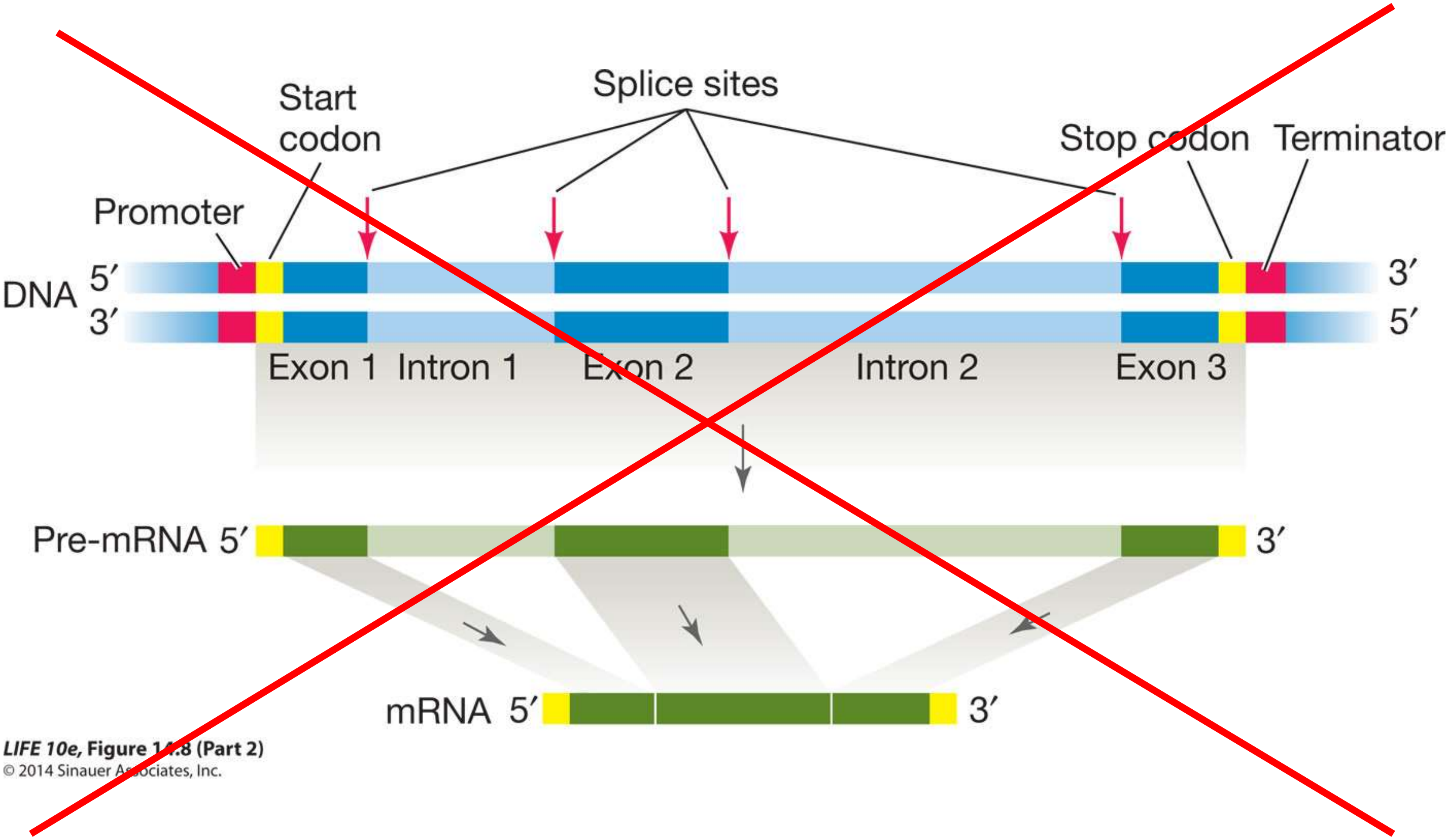
Since there are two strands in DNA, there are *six* possible reading frames in a piece of DNA (three in each direction):

```
3      A  Q  A  E  *  R  R  G  V  F  I  I  *  G  R  C  I
2     C  P  S  *  I  A  *  R  G  F  H  H  L  R  T  M  Y
1    M  P  K  L  N  S  V  E  G  F  S  S  F  E  D  D  V  *
5'  ATGCCCAAGCTGAATAGCGTAGAGGGGTTTTCATCATTTGAGGACGATGTATAA 3'
3'  TACGGGTTCGACTTATCGCATCTCCCCAAAAGTAGTAAACTCCTGCTACATATT 5'
     H  G  L  Q  I  A  Y  L  P  K  *  *  K  L  V  I  Y  L  -1
       G  L  S  F  L  T  S  P  N  E  D  N  S  S  S  T  Y   -2
      A  W  A  S  Y  R  L  P  T  K  M  M  Q  P  R  H  I    -3
```

A reading frame from a start codon to the first stop codon is called an *open* reading frame (underlined above).

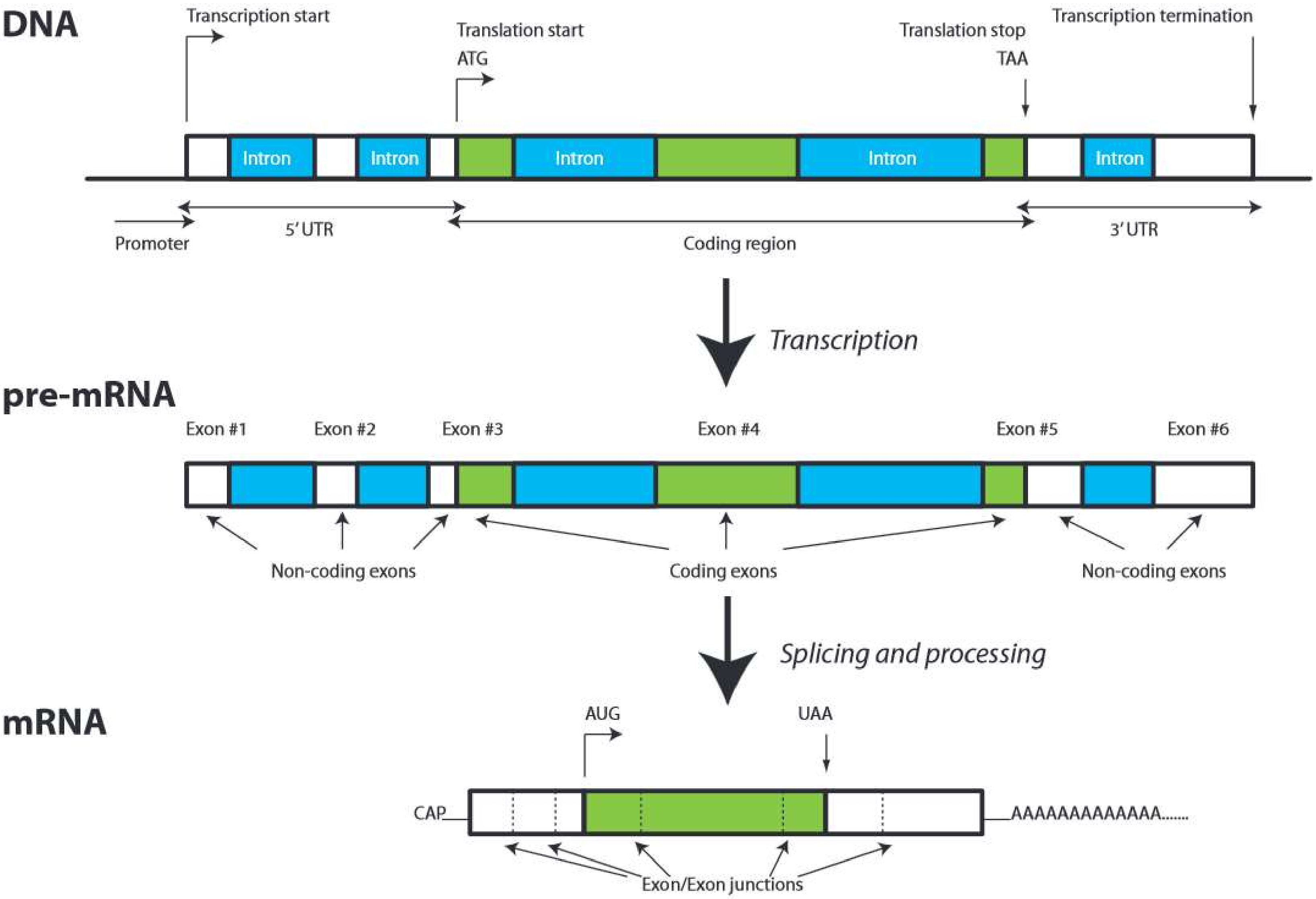
Introns are spliced out

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

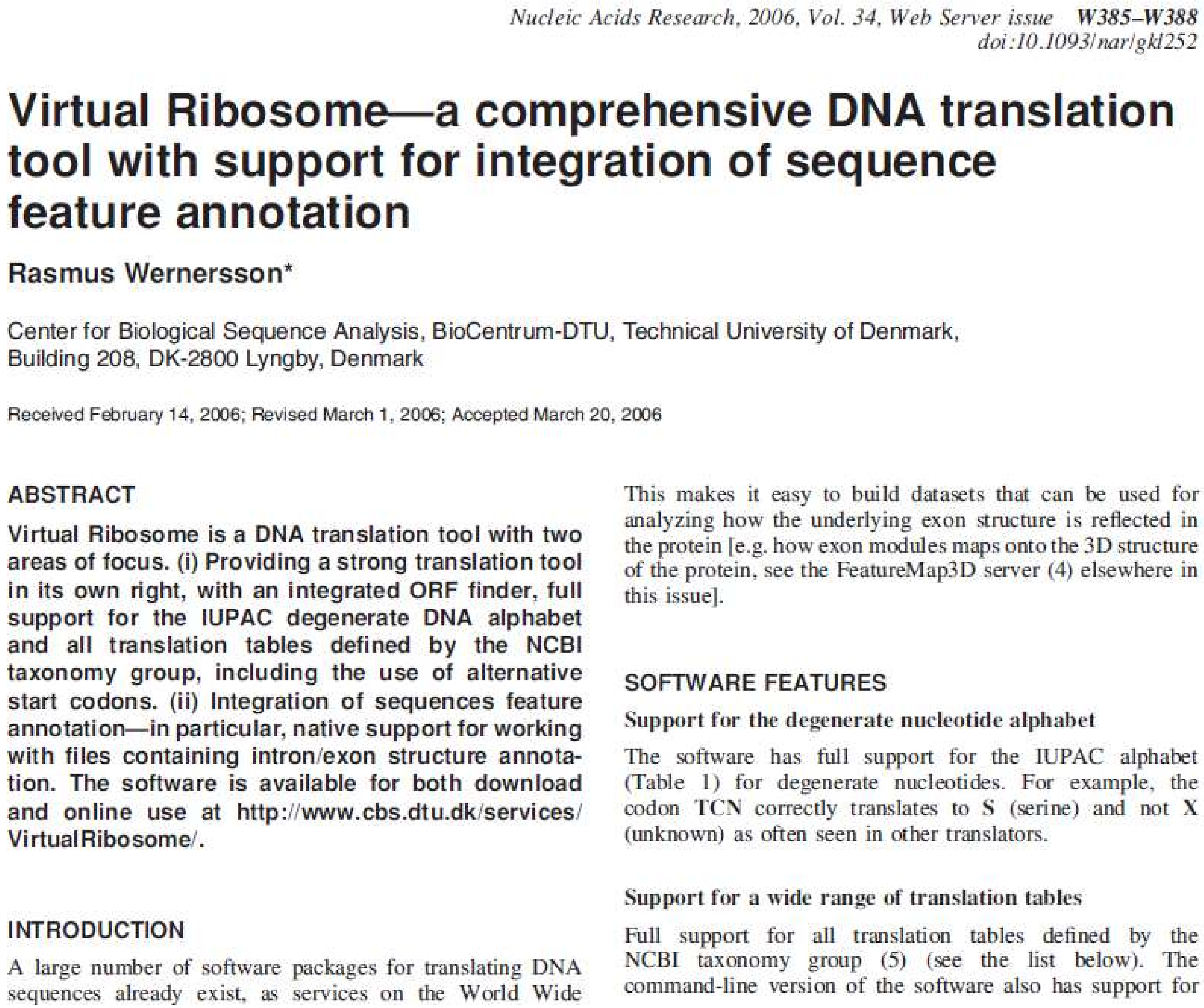


LIFE 10e, Figure 14.8 (Part 2)
© 2014 Sinauer Associates, Inc.

Eukaryotic gene structure



Virtual Ribosome (Curriculum)



Nucleic Acids Research, 2006, Vol. 34, Web Server issue **W385–W388**
doi:10.1093/nar/gkl252

Virtual Ribosome—a comprehensive DNA translation tool with support for integration of sequence feature annotation

Rasmus Wernersson*

Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark,
Building 208, DK-2800 Lyngby, Denmark

Received February 14, 2006; Revised March 1, 2006; Accepted March 20, 2006

ABSTRACT

Virtual Ribosome is a DNA translation tool with two areas of focus. (i) Providing a strong translation tool in its own right, with an integrated ORF finder, full support for the IUPAC degenerate DNA alphabet and all translation tables defined by the NCBI taxonomy group, including the use of alternative start codons. (ii) Integration of sequences feature annotation—in particular, native support for working with files containing intron/exon structure annotation. The software is available for both download and online use at http://www.cbs.dtu.dk/services/VirtualRibosome/.

INTRODUCTION

A large number of software packages for translating DNA sequences already exist, as services on the World Wide

This makes it easy to build datasets that can be used for analyzing how the underlying exon structure is reflected in the protein [e.g. how exon modules maps onto the 3D structure of the protein, see the FeatureMap3D server (4) elsewhere in this issue].

SOFTWARE FEATURES

Support for the degenerate nucleotide alphabet

The software has full support for the IUPAC alphabet (Table 1) for degenerate nucleotides. For example, the codon **TCN** correctly translates to **S** (serine) and not **X** (unknown) as often seen in other translators.

Support for a wide range of translation tables

Full support for all translation tables defined by the NCBI taxonomy group (5) (see the list below). The command-line version of the software also has support for