UniProt

Website and service development updates

Klemens Pichler
We Aim To Provide...

... the highest quality, most comprehensive and most thoroughly annotated protein resource:

- **Detailed information on protein function**, interactions, pathways etc.
- **Sequences**, including isoforms, disease variants and PTMs
- **Stable identifiers (accessions)**

Variations and mutations link to disease and phenotype, or affect drug metabolism.
UniProt (Knowledgebase)

- millions of protein sequences
- experimental information extracted from scientific literature
- provision of complete and up to date Reference proteomes
- integration of large-scale genomics and proteomics experiments
- interoperability between data and services in the biological, medical, translational and clinical domains
Integrating proteomics data

Towards the identification and characterization of cellular proteins that are present, absent, or altered under certain environmental or (patho)physiological conditions
Annotations – Product-specific

**Function**

*Isoform 1:* Involved in lipid metabolism by regulating the size and distribution of lipid droplets. (1 Publication)

**Subcellular location**

*Isoform 1:*
- Endoplasmic reticulum membrane (1 Publication)
- Peripheral membrane protein (1 Publication)
- Nucleus membrane (1 Publication)
- Lipid droplet (1 Publication)
- Cytoplasm > cytoskeleton (2 Publications)
- Endosome (1 Publication)

*Note:* Forms a intramembrane hairpin-like structure in the membrane (PubMed:20200447). Recruited to nuclear membrane by IST1 during late anaphase (PubMed:26040712). (1 Publication)

*Isoform 3:*
- Cytoplasm (1 Publication)
- Endosome (2 Publications)
- Nucleus membrane (2 Publications)

*Note:* Constitutes the main endosomal form (PubMed:19000169). Recruited to nuclear membrane by IST1 during late anaphase (PubMed:26040712). (2 Publications)

Q9UBP0 (SPAST_HUMAN)
## Annotations – PTMs

### Amino acid modifications

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified residue i</td>
<td>1 – 1</td>
<td></td>
<td>N-acetylmethionine</td>
<td></td>
</tr>
<tr>
<td>Modified residue i</td>
<td>36 – 36</td>
<td></td>
<td>Phosphoserine; in isoform p66Shc</td>
<td></td>
</tr>
<tr>
<td>Modified residue i</td>
<td>139 – 139</td>
<td></td>
<td>Phosphoserine</td>
<td></td>
</tr>
<tr>
<td>Modified residue i</td>
<td>154 – 154</td>
<td></td>
<td>Phosphoserine</td>
<td></td>
</tr>
<tr>
<td>Modified residue i</td>
<td>349 – 349</td>
<td></td>
<td>Phosphoserine</td>
<td></td>
</tr>
</tbody>
</table>

- **Added based on publication**
- **Added based on sequence similarity**
- **Added based on proteomics data**
Annotations – PTMs

<table>
<thead>
<tr>
<th>Feature key</th>
<th>Position(s)</th>
<th>Length</th>
<th>Description</th>
<th>Graphical view</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modified residue</td>
<td>293 – 293</td>
<td>1</td>
<td>Phosphoserine</td>
<td></td>
</tr>
<tr>
<td>Modified residue</td>
<td>298 – 298</td>
<td>1</td>
<td>Phosphoserine</td>
<td></td>
</tr>
<tr>
<td>Modified residue</td>
<td>304 – 304</td>
<td>1</td>
<td>Phosphoserine; by CDK1 and CDK2; in isoform 2</td>
<td></td>
</tr>
</tbody>
</table>

Added based on proteomics data; isoform-specific
Large-scale proteomics data in UniProt

UniProt reference proteome sequence collection
species-specific (isoforms always included, variants as optional)

- *In-silico* digestion (any cleaving agent):
  loose/strict trypsin, loose/strict Lys-C, Glu-C (aka V8-DE), chymotrypsin, Asp-N and Lys-N
  two missed cleavages
  with and without initiator methionine cleavage

- Peptide filtering: removal of 6 AA or shorter, removal of X-, B- or Z-containing
- Peptide gene-centric unicity evaluation and distinction between unique and non-unique peptides
- The information of which UniProt sequence produced which peptide, is retained

MS-proteomics repositories
*filtered* lists of peptides
species-specific

Annotation of UniProt protein sequences
Large-scale proteomics data in UniProt

P05067 (A4_HUMAN)
Large scale proteomics - human

UniProt human reference proteome 92578 sequences for which at least one unique peptide from at least one repository is mapped

- 80% with evidence
- 20% without evidence

Some numbers
Human 2016_06 had 20191 gene groups containing 92578 UniProt accession numbers of which 79.70% (75782) have been mapped to peptides in public repositories
## Large scale proteomics

<table>
<thead>
<tr>
<th>Species (UniProt release 2016_07)</th>
<th>UPID</th>
<th>TaxID</th>
<th>Sequences</th>
<th>SeqWithUniqueEvidence</th>
<th>% SeqWithUnEv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caenorhabditis elegans</td>
<td>UP000001940</td>
<td>6239</td>
<td>27945</td>
<td>15627</td>
<td>55.92</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>UP000000803</td>
<td>7227</td>
<td>23300</td>
<td>10044</td>
<td>43.11</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>UP000000437</td>
<td>7955</td>
<td>43657</td>
<td>13507</td>
<td>30.94</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>UP000005640</td>
<td>9606</td>
<td>92578</td>
<td>73782</td>
<td>79.70</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>UP000008227</td>
<td>9823</td>
<td>26238</td>
<td>7458</td>
<td>28.42</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>UP000009136</td>
<td>9913</td>
<td>24582</td>
<td>1182</td>
<td>4.81</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>UP000000589</td>
<td>10090</td>
<td>59521</td>
<td>41237</td>
<td>69.28</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>UP00002494</td>
<td>10116</td>
<td>31664</td>
<td>12515</td>
<td>39.52</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe (strain 972 / ATCC 24843)</td>
<td>UP00002485</td>
<td>284812</td>
<td>5128</td>
<td>4311</td>
<td>84.07</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (strain ATCC 204508 / S288c)</td>
<td>UP00002311</td>
<td>559292</td>
<td>6749</td>
<td>4813</td>
<td>71.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Species (UniProt release 2016_07)</th>
<th>UPID</th>
<th>TaxID</th>
<th>MatchedUniquePeps</th>
<th>MatchedNonUniquePeps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caenorhabditis elegans</td>
<td>UP000001940</td>
<td>6239</td>
<td>151924</td>
<td>4134</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>UP000000803</td>
<td>7227</td>
<td>34419</td>
<td>546</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>UP000000437</td>
<td>7955</td>
<td>58860</td>
<td>5414</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>UP000005640</td>
<td>9606</td>
<td>614991</td>
<td>30983</td>
</tr>
<tr>
<td>Sus scrofa</td>
<td>UP000008227</td>
<td>9823</td>
<td>49764</td>
<td>6686</td>
</tr>
<tr>
<td>Bos taurus</td>
<td>UP000009136</td>
<td>9913</td>
<td>2523</td>
<td>375</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>UP000000589</td>
<td>10090</td>
<td>318054</td>
<td>13582</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>UP00002494</td>
<td>10116</td>
<td>135454</td>
<td>10962</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe (strain 972 / ATCC 24843)</td>
<td>UP000002485</td>
<td>284812</td>
<td>60345</td>
<td>989</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (strain ATCC 204508 / S288c)</td>
<td>UP000002311</td>
<td>559292</td>
<td>83292</td>
<td>2963</td>
</tr>
</tbody>
</table>
Peptide search tool
UniProt tools

BLAST
Align
Retrieve/ID mapping
Peptide search

BLAST Sequence search
Multiple sequence alignment
Batch retrieval/ ID mapping
Peptide search

NEW!
Peptide search

- Finds all UniProtKB sequences that exactly match a peptide sequence
  - Allows treating isoleucine and leucine as equivalent in the search

- Allows searching for multiple peptides at once

- Allows restriction to a taxon (optional)
Peptide search

How to use this tool

The Peptide search tool finds all UniProtKB sequences that exactly match a query peptide sequence.

1. Enter peptide sequences, separated by spaces or new lines, using single letter amino acid codes (3 or more), into the form field, for example:
   - RVLSLGR
   - LDEAEFEVK
2. Optionally, enter a taxon to restrict the search space.
3. Click the Run Peptide search button.

Search for 1 or many peptides at once

Restrict search to a taxon (optional): e.g. Homo sapiens

- Treat Isoleucine and Leucine as equivalent.
- Run Peptide search in a separate window.

Clear  Run Peptide search
Peptide search

Positions of hits are given

Job is stored for 7 days

Usual set of filters is available
Publications in UniProtKB
Annotation from publications

Unreviewed UniProtKB entries contain publications documenting sequence origin and mapped 3D structures

Reviewed UniProtKB entries contain publications for functional characterization, sequence origin and mapped 3D structures
Publications in the entry view

Attached to individual annotations as part of evidence tags via dedicated publication view
New representation of publications...

... incorporating filters and access to mapped publications.

Publications related to Q7Z2E3 - APTX_HUMAN

1. "Novel splice variants increase molecular diversity of aprataxin, the gene responsible for early-onset ataxia with ocular motor apraxia and hypoaalbuminemia."
   Hirano M., Nishikawa T., Kariya S., Furiya Y., Kawahara M., Uenc S.
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS 7; 12 AND 13), ALTERNATIVE SPlicing, SUBCELLULAR LOCATION.
   Category: Subcellular Location, Sequences.
   Source: UniProtKB/Swiss-Prot (reviewed).
   This publication is cited by 1 other entry.

   Huang C.-H.
   Submitted (JUN-2001) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
   Category: Sequences.
   Source: UniProtKB/Swiss-Prot (reviewed).

3. Chen Y., Huang C.-H.
   Submitted (DEC-2002) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS 1; 2; 3; 4; 5; 6; 7; 8 AND 10).
   Category: Sequences.
   Tissue: Hypothalamus, Kidney, Lung adenocarcinoma, Lymphoma, Melanoma, Muscle, Retinoblastoma, Skin and Testis.
   Source: UniProtKB/Swiss-Prot (reviewed).

4. "Complete sequencing and characterization of 21,243 full-length human cDNAs."
   Shibahara T., Tanaka T., Ishii S., Sugano S.
   Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORMS 2 AND 11).
   Category: Sequences.
   Tissue: Colon.
   Source: UniProtKB/Swiss-Prot (reviewed).
   This publication is cited by 12919 other entries.

5. "The full-ORF clone resource of the German cDNA consortium."
   Koehler K., Ottenwaelder B., Pousta A., Wiemann S., Schupp I.
   BMC Genomics 8:399-399(2007) [PubMed] [Europe PMC] [Abstract]
   Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORM 9).
Two sources for publications

Publications related to Q7Z2E3 - APTX_HUMAN

1. "Novel splice variants increase molecular diversity of aprataxin, the gene responsible for early-onset ataxia with ocular motor apraxia and hypoalbuminemia."
   Hirano M., Nishiwaki T., Karlya S., Furiya Y., Kawahara M., Ueno S.
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS 7; 12 AND 13), ALTERNATIVE SPlicing, SUBCELLULAR LOCATION.
   Category: Subcellular Location, Sequences.
   Source: UniProtKB/Swiss-Prot (reviewed).
   This publication is cited by 1 other entry.

   Huang C.-H.
   Submitted (JUN-2001) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
   Category: Sequences.
   Source: UniProtKB/Swiss-Prot (reviewed).

3. Chen Y., Huang C.-H.
   Submitted (DEC-2002) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORMS 1; 2; 3; 4; 5; 6; 7; 8 AND 10).
   Category: Sequences.
   Tissue: Hypothalamus, Kidney, Lung adenocarcinoma, Lymphoma, Melanoma, Muscle, Retinoblastoma, Skin and Testis.
   Source: UniProtKB/Swiss-Prot (reviewed).

4. "Complete sequencing and characterization of 21,243 full-length human cDNAs."
   Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORMS 2 AND 11).
   Category: Sequences.
   Tissue: Colon.
   Source: UniProtKB/Swiss-Prot (reviewed).
   This publication is cited by 12919 other entries.

5. "The full-ORF clone resource of the German cDNA consortium."
   BMC Genomics 8:399-399(2007) [PubMed] [Europe PMC] [Abstract]
   Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORM 9).
Computationally mapped publications

- imported from curated resources:
  - e.g. PDB, Intact, FlyBase, WormBase, MGI, RGD
  - categorized based on source, e.g. PDB -> interaction

- collated by text mining:
  - PgenN (some plant species)
  - PubTator_TmVAR (proteins/variants/disease)
Filter publications by topic

Publications related to Q7Z2E3 - APTX_HUMAN

21. "CK2 phosphorylation-dependent interaction between aprataxin and MDC1 in the DNA damage response."
   Nucleic Acids Res. 38:1489-1503(2010) [PubMed] [Europe PMC] [Abstract]
   Cited for: X-RAY CRYSTALLOGRAPHY (1.65 ANGSTROMS) OF 15-116, SUBCELLULAR LOCATION, INTERACTION WITH MDC1, MUTAGENESIS OF ARG-43 AND LYS-52.
   Category: Subcellular Location, Pathology & Biotech, Interaction, Structure.
   Source: UniProtKB/Swiss-Prot [reviewed].
   This publication is cited by 1 and mapped to 1 other entries.

33. "Aprataxin resolves adenylated RNA-DNA junctions to maintain genome integrity."
   Tumbale P., Williams J.S., Schellenberg M.I., Kunkel T.A., Williams R.S.
   Nature 506:111-115(2014) [PubMed] [Europe PMC] [Abstract]
   Category: Structure.
   Source: PDB:4NDF, PDB:4NDG, PDB:4NDH, PDB:4NDL.
   This publication is mapped to 4 other entries.
The Team

**PIs:** Alex Bateman, Cathy Wu, Ioannis Xenarios

**Key staff:** Cecilia Arighi (Curation), Lydie Bougueleret (Co-DIRECTION), Alan Bridge (Content), Hongzhan Huang (Development), Michele Magrane (Curation), Maria Martin (Development), Peter McGarvey (Content), Darren Natale (Content), Claire O'Donovan (Content), Sylvain Poux (Curation), Manuela Pruess (Coordination), Nicole Redaschi (Development)


**European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK**

**Protein Information Resource (PIR), Washington DC and Delaware, USA**

**SIB Swiss Institute of Bioinformatics (SIB), Geneva, Switzerland**