Expert curation in UniProtKB/Swiss-Prot

sylvain.poux@sib-sib.ch
Head of annotation department
Swiss-Prot group
SIB Swiss Institute of Bioinformatics
San Diego, January 2016

Expert curation of plant proteins in UniProtKB

The Universal Protein resource

- comprehensive
- high quality
- and freely accessible

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

data

protein sequence

knowledge

functional information

www.uniprot.org

UniProtKB, the Knowledge base component of UniProt

Where data

UniProtKB, the Knowledge base component of UniProt

Where data becomes structured knowledge
Expert curation of plant proteins in UniProtKB

UniProtKB/TrEMBL - Automatic annotation

- Protein sequence: The quality of the protein sequences is dependent on the information provided by the submitter of the original nucleotide entry (CDS) or of the gene prediction pipeline (i.e. Ensembl).
- Biological information: 100% identical sequences (same length, same organism) are combined into a single record.

UniProtKB/Swiss-Prot – Expert curation

- Protein sequence: (combine together available CDS, annotate sequence discrepancies, report sequencing mistakes...)
- Biological information: (sequence analysis, extract literature information, ortholog data propagation, ...)

Biological information

- Sources of annotation: Provided by the submitter (EMBL, PDB, Ensembl...)
- Imported (GO annotation from TAIR)
- Automatic annotation from rules

What do UniProtKB/Swiss-Prot biocurators do?

They read publications and extract the KNOWLEDGE that will help explain the function of a protein.

From January 2015 to January 2016
More than 8,000 articles have been curated by UniProtKB/Swiss-Prot biocurators

What is inside a UniProtKB/Swiss-Prot protein entry?

Sequence

- Sequence features

UniProtKB release 2015_12

37,193 plant protein sequences manually curated in UniProtKB/Swiss-Prot, including 14,349 from Arabidopsis
Expert curation of plant proteins in UniProtKB

- http://www.uniprot.org/program/plants/
- Program managed by Michel Schneider
- Composed of 4 curators
- Focuses on the manual annotation of plant-specific proteins and plant-specific protein families.
- Organism priorities: A.thaliana (dicot) and O.sativa subsp japonica (monocot)
- Focus on proteins with functional characterization

---

The case of miPEP165a and miPEP171b

**LETTER**

Primary transcripts of microRNAs encode regulatory peptides

(Nature 520:90-93(2015))

- New protein sequences, not predicted by gene prediction programs
- No AGI number present

---

The case of miPEP165a and miPEP171b

AGI number submitted to and validated by Araport

---

Annotations – Evidence attributions

Q05W2 (CHR35_ARATH)

**Function**

- GO: Biological process
  - Cellular response to external stimulus
  - IAA metabolism
Expert curation of plant proteins in UniProtKB

Status of the complete proteome of A.thaliana

- 14,349 entries as well as 1,954 manually reviewed alternative products
- Most protein-coding genes in Arabidopsis functionally characterized are present in UniProtKB/Swiss-Prot

Unreviewed (TrEMBL) - Computationally analyzed

- 17,202 additional entries, corresponding to uncharacterized proteins and to redundant entries, are available in the TrEMBL section of UniProtKB and can be retrieved as part of the complete Arabidopsis thaliana proteome

Curation is not limited to A.thaliana and O.sativa proteins
- Characterized proteins from other plant organisms are also curated

Morphine biosynthesis proteins

UniProtKB - P0DK7 (STORR_PAPSO)

Function:

- Curation is not limited to A.thaliana and O.sativa proteins
- Characterized proteins from other plant organisms are also curated

Morphine biosynthesis pathway

Proteomics in UniProtKB

- Methods and technologies evolve
- Should limit the number of false positives
- Established our own pipeline for annotation of proteomics data

Pipeline for integration of high-throughput proteomics data
Expert curation of plant proteins in UniProtKB

Proteomics in UniProtKB

Collect all sequences from complete proteomes
Canonical, isoforms, processed sequences, variants

Compute all possible peptides (7-50 AA)

Selection of the paper

Annotation of UniProtKB

A. thaliana proteomics in UniProtKB

- **Source:** ARATH phosphoproteome
  - Identification by MS:
    - 4,545 peptides
    - 416 phosphopeptides (345 phosphoproteins)
- **Processing by our pipeline**
  - Apply of stringent filtering rules:
    - **Peptide validation**:
      - Charge=2:
        - Mascot: Identity score > 35
        - Sequest: xcorr > 3 OR Sequest: dcn > 0.2
        - X!tandem: expect >= 3
    - **Charge=3**:
      - Mascot: Identity score > 40
      - Sequest: xcorr > 3.5 AND Sequest: dcn > 0.2
      - X!tandem: expect >= 3
    - **PTM validation**:
      - Only manually confirmed by authors
- **Filtered data**
  - 5,048 filtered peptides
    - Including 168 certified phosphopeptides (146 phosphoproteins)
- Integration into UniProt entries

UniProtKB/Swiss-Prot

- is an expertly curated protein sequence knowledgebase
- contains functional information
- serves as central hub for biomolecular information
- is a freely available service for the scientific community
- collaborates with Model Organism Databases (MODs) and the scientific community
- participates to the curation of GO terms
- keeps the scientific content of UniProtKB up-to-date
- integration of high-throughput proteomics data

Summary

help@uniprot.org

Thank you for your attention and thanks to the UniProt teams at:

- SIB
- EBI
- PIR

Funding agencies:

- National Institutes of Health (NIH) (NHGRI and NIGMS)
- Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI)
- The European Commission