Integration of proteomics data into UniProtKB

Background
- The amount of publicly available data in mass spectrometry (MS) proteomics repositories like PRIDE, PeptideAtlas, MaxQB, EPD, CTDP, ProteomicsDB, GPMDB, HPM and others, as well as data from third-party global PRIDE reprocessing, is growing at a fast rate.
- The Universal Protein Resource (UniProt, http://www.uniprot.org/) is integrating proteomics data from all available public sources to provide comprehensive protein data sets with high quality experimentally proven evidence.

UniProt protein existence (PE) values (http://www.uniprot.org/manual/protein_existence)

<table>
<thead>
<tr>
<th>PE value</th>
<th>Nr of sequences (canonic/alt)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Evidence at protein level</td>
<td>2350</td>
<td>0.25</td>
</tr>
<tr>
<td>2. Evidence at transcript level</td>
<td>1354972</td>
<td>1.73</td>
</tr>
<tr>
<td>3. Inferred from homology</td>
<td>13635703</td>
<td>22.39</td>
</tr>
<tr>
<td>4. Predicted</td>
<td>4859131</td>
<td>75.72</td>
</tr>
<tr>
<td>5. Uncertain</td>
<td>1560</td>
<td>0.00</td>
</tr>
</tbody>
</table>

UniProt release 2016_02

MS-proteomics repositories filtered lists of peptides species-specific

PepTides

Peptides

Exact match

UniProtKB/TrEMBL accessions with unique peptides are assigned a keyword to trigger PE updated to 1.

Unique/none/unique peptides/source(s) are provided on the ftp for both UniProtKB/Swiss-Prot and UniProtKB/TrEMBL.

Gene-centric unicity: a peptide is considered unique if it belongs to only one specific gene group. Each group is constituted by one or more UniProtKB sequences. Gene groups are an underlying feature of the UniProtKB reference proteomes. Representative symbols for each gene group are provided in the ftp output files.

Proteomics mapping files are publicly available since UniProt release 2015_03 (ftp://uniprot.ebi.ac.uk/help/2015/03/releases) from a dedicated section on the Downloads page of the UniProt website (http://www.uniprot.org/downloads), which points to the UniProt ftp (ftp://ftp.uniprot.org/pdb databases/uniprot_released/knowledgebase/proteomics_mapping/nov0104.txt) and a readme file (ftp://ftp.uniprot.org/pdb databases/uniprot_released/knowledgebase/proteomics_mapping/README) are also provided. Mappings are currently available for ten species coming from three (since UniProt release 2016_03) MS proteomics repositories. More species will be added as soon as additional MS proteomics repositories will be integrated.

Next developments
- All the proteomics data incorporated so far are bottom-up data; the Consortium for Top Down Proteomics has started providing top-down. Cross references to CTDP are in place since UniProt release 2016_03.
- Provide a visual representation of the mapped peptides (unique/non-unique) via the UniProt feature viewer (http://www.uniprot.org/help/2016/02/17/release available since UniProt release 2016_02.

Contact: help@uniprot.org

UniProt postdocs available at: http://www.ebi.ac.uk/uniprot/postdocs

UniProt is mainly supported by the National Institutes of Health (NIH) grant U54HG007932. Additional support for the EMBL-EBI’s involvement in UniProt comes from European Molecular Biology Laboratory (EMBL), the British Heart Foundation (BHFRG15031132), the Parkinson’s Disease United Kingdom PTUKN CLG grant G0707, and the NIH-C2 grant U54HG007236. UniProt activities at the SRB are additionally supported by the State Federal Government through the State Committee for Education, Research and Innovation SERI. PDB’s UniProt activities are also supported by the NIH grants R01DK095894, RO1HL101702, and FO20M103466, and the National Science Foundation (NSF) grants DBI-0652420.