UniProt -
The Universal Protein Resource

Claire O’Donovan
Pre-UniProt

- **Swiss-Prot**: created in July 1986; since 1987, a collaboration of the SIB and the EMBL/EBI;

- **TrEMBL**: created at the EBI in November 1996 as a computer-annotated protein sequence database supplementing Swiss-Prot. It was introduced to deal with the increased data flow from genome projects.
The UniProt timeline

- Awarded to EBI, SIB, and PIR by NIH
- Run time 9/02-8/05
- ~16 million USD intended to replace Swiss-Prot license fees and previous PIR funding
UniProt Consortium

EMBL – EBI
European Bioinformatics Institute

UniProt
Universal Protein Resource
UniProt Consortium activities
The three-layered approach

- **The UniProt Archive (UniParc)**
  - ✔ UniProtKB + all other protein sequences publicly available
  - ✔ Completeness

- **The UniProt Reference Clusters (UniRef)**
  - ✔ Non-redundant views of UniProtKB + selected UniParc sets
  - ✔ Speed

- **The UniProt Knowledgebase (UniProtKB)**
  - ✔ Central database of annotated protein sequences and functional information
  - ✔ UniProtKB/Swiss-Prot + UniProtKB/TrEMBL
The three layer approach

Interrelationship between the UniProt Databases
UniProt Archive

- UniParc is a non-redundant archive of protein sequences from the public databases
- It contains only protein sequences (no annotations)
- It provides cross-references to the source databases
UniProt Archive: Principles

- UniParc is non-redundant
  - Each unique protein sequence is stored only once and is assigned a unique stable UniParc identifier (e.g. UPI0000000356)
- UniParc provides cross-references to the original source: active or retired
- UniParc provides sequence versions.
UniProt Reference Clusters

Principles

- It provides non-redundant reference data collections
- It allows faster and more informative sequence similarity searches
- It includes the UniProtKB and some data from UniParc
- It merges across different species
UniProt Reference Clusters
Principles

- UniRef100
  - It merges identical sequences and subfragments
- UniRef90
  - Size reduction of 40%
- UniRef50
  - Size reduction of 65%
<table>
<thead>
<tr>
<th>UniProtKB/Swiss-Prot</th>
<th>UniProtKB/TrEMBL</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Non-redundant</td>
<td>- Translations of CDS in</td>
</tr>
<tr>
<td>- High level of integration</td>
<td>EMBL/GenBank/DDBJ</td>
</tr>
<tr>
<td>- High level of manual curation</td>
<td>- Automatic annotation</td>
</tr>
<tr>
<td>- Contains 241,242 entries</td>
<td>- Contains 3,313,265 entries</td>
</tr>
</tbody>
</table>
UniProtKB/TrEMBL

- Automatically generated in a biweekly cycle from the data present in EMBL/GenBank/DDBJ and some other sources such as TAIR/SGD
- Exclusions: pseudogenes, synthetic, immunoglobulins, patents, small sequences <8
- /product, /gene, /locus_tag
- RefSeq and Ensembl
UniProtKB/TrEMBL

- Proteome annotation
- Cross-references to other databases
- Addition of relevant publications (eg PDB)
- Redundancy
- Automatic annotation
- Future plans for manual annotation eg human proteome project
Capturing the correct sequence

- Archive collections
- Each sequence report stored in its own entry

- Merging at 100% identity
- Still some redundancy
Sequence similarity searches

- Identify potential merge candidates
- Identify similar already curated entries
Sequence comparison

- Sequence alignments
- Identification of sequence differences
- Helps in identifying underlying causes
Causes of sequence differences

- Polymorphisms, disease variants
- Splice variants
- Sequencing errors
- Incorrect predictions
Literature curation

- 1741 different journals cited in Swiss-Prot
- Total of 383,401 references
- Average of 2 references per entry
| **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 physiological events. |
| **CATALYTIC ACTIVITY** | Specific cleavage of Arg-I-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 (LRP). 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. Binds LRP1B; binding is followed by internalization and degradation. |
| **SUBCELLULAR LOCATION** | Secreted; extracellular. |
| **ALTERNATIVE PRODUCTS** | Alternative splicing; 2 named isoforms. [Display all isoform sequences in Fasta format] |
| **Name** | 1 |
| **Synonyms** | Long |
| **IsoformId** | P00750-1 |
| **Sequence** | This is the isoform sequence displayed in this entry. |
| **Name** | 2 |
| **Synonyms** | Short |
| **IsoformId** | P00750-2 |
| **Sequence** | VSP_005411, VSP_005412 |
| **Note** | May be produced at very low levels due to a premature stop codon in the mRNA, leading to nonsense-mediated mRNA decay. |
| **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. |
| **PTM** | N-glycosylation of Asn-152; the bound oligomannoside 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 initiate fibrinolysis. |
| **SIMILARITY** | Contains 1 EGF-like domain. |
| **SIMILARITY** | Contains 1 fibronectin type-I domain. |
| **SIMILARITY** | Contains 2 kringle domains. |
| **SIMILARITY** | Contains 1 peptidase S1 domain. |
| **SIMILARITY** | Belongs to the peptidase S1 family. |
Sequence analysis

- Range of sequence analysis tools used to predict important sequence features
- Use of most appropriate programs
- Development of new predictive methods
Evidence attribution

- System which allows linking of all information in an entry to its original source.

- Allows users:
  - to trace origin of all data
  - to differentiate easily between literature-derived and computational data
  - to assess data reliability
UniProtKB curation group

14 curators

24 curators

2 curators
EBI curation projects

- Submissions
- Journal scanning
- Species-specific curation
  - human, mouse, rat, C.elegans, Drosophila, Xenopus, zebrafish, S.cerevisiae, S.pombe
- Protein family curation
  - kinases, keratins
- UniProtKB-MSD collaboration
- PTM standardisation
Some future curation plans

- Improvements to SPIN
- Extension of evidence attribution system to Swiss-Prot
- New annotation projects
- Community participation
- Further database collaborations
UniProt distribution

- Biweekly distribution
- Website access www.uniprot.org
- FTP access
- DVD of UniProtKB (datalib@ebi.ac.uk)
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 (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Reference Clusters (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.
The new UniProt grant timeline

- Second Grant awarded to EBI, SIB, and PIR by NIH

- Run time 9/06-8/09
# Acknowledgements (1)

<table>
<thead>
<tr>
<th>Production:</th>
<th>Proteomes:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel Barrell</td>
<td>Alan Horne</td>
</tr>
<tr>
<td>Renato Golin</td>
<td>Paul Kersey</td>
</tr>
<tr>
<td>Alexander Fedetov</td>
<td></td>
</tr>
<tr>
<td>Maria Jesus Martin</td>
<td></td>
</tr>
<tr>
<td>Patricia Monteiro</td>
<td></td>
</tr>
<tr>
<td>Claire O’Donovan</td>
<td></td>
</tr>
<tr>
<td>Mark Rijnbeek</td>
<td></td>
</tr>
<tr>
<td>UniParc/UniSave:</td>
<td>AutomaticAnnotation</td>
</tr>
<tr>
<td>Quan Lin</td>
<td>/Kraken/Website/XML:</td>
</tr>
<tr>
<td>Andrey Sitnov</td>
<td>Michael Kleen</td>
</tr>
<tr>
<td>Rasko Leinonen</td>
<td>Ernst Kretschmann</td>
</tr>
<tr>
<td></td>
<td>John O’Rourke</td>
</tr>
<tr>
<td></td>
<td>Sam Patient</td>
</tr>
<tr>
<td></td>
<td>Emilio Salazar</td>
</tr>
<tr>
<td></td>
<td>Natalyia Skylar</td>
</tr>
<tr>
<td></td>
<td>Dani Wieser</td>
</tr>
</tbody>
</table>
Acknowledgements (2)

EBI curators:
- Michele Magrane (Annotation coordinator / Mouse)
- Yasmin Alam (Keratins)
- Paul Browne (Journal scan)
- Wei Mun Chan (Human)
- Ruth Eberhardt (Submissions)
- Rebecca Foulger (Xenopus)
- Gill Fraser (Zebrafish)
- Gabriella Frigerio (Rat)
- John Garavelli (PTMs)
- Jules Jacobsen (Structural data)
- Kati Laiho (Fungi)
- Claire O’Donovan (Quality control, data integration)
- Sandra Orchard (Kinases)
- Eleanor Whitfield (C.elegans, Drosophila)

- SIB Group
- PIR Group
- Rolf Apweiler