Welcome - webinar instructions

• GoToTraining works best in **Chrome** or **IE** — avoid Firefox due to audio issues with Macs

• To access the full features of GoToTraining, use the desktop version by clicking on the flower symbol and selecting “**switch to desktop version**”

• **All microphones will be muted** whilst the trainer is speaking

• If you have a question please use the **chat box** at the bottom of the GoToTraining box

---

**EMBL-EBI**
Automatic annotation systems in UniProteomics

Alexandre Renaux
Klemens Pichler

help@uniprot.org
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)**
Pls: Alex Bateman, Cathy Wu, Ioannis Xenarios

Key staff: Alan Bridge (Content), Cecilia Arighi (Curation), Darren Natale (Content), Hongzhan Huang (Development), Lydie Bougueleret (Co-direction), Manuela Pruess (Coordination), Maria Martin (Development), Michele Magrane (Curation), Nicole Redaschi (Development), Peter McGarvey (Content), Sandra Orchard (Content), Sylvain Poux (Curation)


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
UniProt funding

UniProt is supported by the **National Institutes of Health (NIH)**, National Human Genome Research Institute (NHGRI) and National Institute of General Medical Sciences (NIGMS) grant U41HG007822.

Additional support for the **EMBL-EBI**'s involvement in UniProt comes from European Molecular Biology Laboratory (EMBL), the British Heart Foundation (BHF) (RG/13/5/30112), the Parkinson's Disease United Kingdom (PDUK) GO grant G-1307, and the NIH GO grant U41HG02273.

UniProt activities at the **SIB** are additionally supported by the **Swiss Federal Government** through the State Secretariat for Education, Research and Innovation SERI.

**PIR**'s UniProt activities are also supported by the NIH grants R01GM080646, G08LM010720, and P20GM103446, and the **National Science Foundation (NSF)** grant DBI-1062520.
Resources we provide and their links

UniProtKB
Protein Knowledgebase

- Swiss-Prot
  Manually annotated and reviewed
- Expert manual annotation
- TrEMBL
  Automatically annotated and not reviewed

External Sources
- EMBL/GenBank/
  DDBJ, Ensembl, VEGA,
  RefSeq, PDB,
  MODs, other sequence
  resources

UniParc
Sequence archive

Proteomes
Protein sets expressed by organisms

UniRef
Sequence clusters
- UniRef100
- UniRef90
- UniRef50

Sequence archive of new, revised
and obsolete sequences
The two datasets in UniProtKB

- **UniProtKB**
  - **Swiss-Prot**
    - ~0.5 million proteins
    - Manually annotated and reviewed.
  - **TrEMBL**
    - ~100 million proteins
    - Automatically annotated and not reviewed.

- **expert-curated information on all aspects of a protein, e.g.**:
  - function
  - residue-specific data (e.g. active sites)

- protein isoforms of one gene in **one** UniProtKB record (per species)

- protein isoforms of one gene in **different** UniProtKB records (per species)

- mapped experimental sequence features (from protein 3D structures)

- annotations from rule systems (incl. expert-curated rules)
Expert curation for UniProtKB entries

nucleotide seqs
(large/small scale)

select

unreviewed

update

Integrate

Reviewed

Test against automatic annotation pipelines

Expert curation

Select

Annotate

Automatic rules

Expert rules
Types of data in a UniProtKB entry

- **Sequences**
  - isoforms
  - variants etc

- **Identifiers**
  - unique, stable

- **Functional annotation**

- **Amino acid-specific data**

- **Names**
  - comprehensive
  - standardized

- **Links to specialized databases**

- **Gene ontology terms**

- **Taxonomy**

---

- **Scientific literature**

- **Sequence curation**

- **Bioinformatics predictions**

- **Family curation**

- **Evidence attribution**

- **Quality control**
Evidence attribution

• describe the source of an annotation
  • annotation = piece of information contained in a UniProtKB entry
• based on Evidence Ontology (ECO)
  • assertion: protein X is alcohol dehydrogenase
  • source data provides supports evaluated by curator
  • ECO:0000269
  • PMID:12345456
  • evidence: experimental
  • method: manual
  • evaluated by curator
Using evidence tags in searches

Evidence filters relevant to Automatic annotation
Automatic annotation systems
Automatic annotation in UniProtKB

- **GROWS FAST**
  - Select
  - Annotate
  - Test against automatic rules

- **TAKES TIME**
  - Expert curation
  - Update

- **LACK OF EXPERIMENTAL DATA**
  - Integrate
  - Reviewed

**Lack of experimental data**

**Time consuming**

**Grows fast**

**Reviewed**

**Automatic**

**Annotation pipelines**

**Expert rules**

**Automatic rules**
Rules – how do they work?

If a protein meets these conditions...

• color: purple
• has four edges: yes

... then these annotations are applied

• is a purple quadrilateral

sequence space (unreviewed)
A fictitious rule

If a protein meets these conditions...

- color: purple
- has four edges: yes

and:

- edges are of same length
- angles are 90 degrees

... then these annotations are applied

- is a purple quadrilateral
- is a square

sequence space (unreviewed)
Components of a rule: conditions

- InterPro signatures
- taxonomy
- sequences features, e.g. length
- proteome features, e.g. *outer membrane*:yes; (bacterial sequences)
InterPro

Functional classification of proteins using predictive models (signatures)

SIGNATURE METHOD
- Profile HMMs
- Profiles
- Patterns
- Composition Prediction

BIOLOGICAL ENTITY
- Homologous Superfamilies
- Domain & Families
- Features & sites
- Intrinsic Disorder

MEMBER DATABASES
- CATH-Gene3D
- SUPERFAMILY
- Pfam
- SMART
- TIGRFAM
- PIRSF
- SFLD
- PANTHER
- HAMAP
- Prosite Profiles
- CDD
- Prodom
- PRINTS
- Prosite Patterns
- MobiDB

INTERPRO

Most relevant to UniProtKB automatic annotation
Components of a rule: conditions

- Sequences: isoforms, variants, etc.
- Functional annotation
- Amino acid-specific data
- Names: comprehensive, standardized
- Identifiers: unique, stable
- Gene ontology terms
- Taxonomy

Links to specialized databases

- CATH-Gene3D
- SUPER-FAMILY
- PRINTS
- Pfam
- SMART
- Prosite Patterns
- TIGRFAM
- PIRSF
- SFLD
- PANTHER
- HAMAP
- Prosite Profiles
- CDD
- Prodom
### Components of a rule: annotations

<table>
<thead>
<tr>
<th>Sequences</th>
<th>Functional annotation</th>
<th>Amino acid-specific data</th>
<th>Names comprehensive standardized</th>
</tr>
</thead>
<tbody>
<tr>
<td>isoforms variants etc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identifiers unique, stable</td>
<td>Links to specialized databases</td>
<td>Gene ontology terms</td>
<td>Taxonomy</td>
</tr>
</tbody>
</table>
Example of a UniRule

UniRule: UR000172789

Source ID: RU362103

If a protein meets these conditions...  

Common conditions
- Matches Pfam signature PF01735
- taxon = Fungi

Special conditions
- Predicted signal

... then these annotations are applied

Protein name
- Recommended name: Lysophospholipase (EC: 3.1.1.5)

Catalytic activity
- 2-lyosphosphatidylcholine + H₂O = glycerophosphocholine + a carboxylate.

Sequence similarities
- Belongs to the lysophospholipase family.

Signal peptide
- (to residues corresponding to positions @NTER@ - @TO@)

Chain
- @CHAIN_NAME@ (to residues corresponding to positions @TO|+1 - @CTER@)

Keywords
- Hydrolase
- Lipid degradation
- Lipid metabolism
- Signal
How to access automatic annotation data?

UniRule annotation

UniProtKB
Protein knowledgebase

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes
Protein sets from fully sequenced genomes

Help
Help pages, FAQs, UniProtKB manual, documents, news archive and Biocuration projects.

Annotation systems
Systems used to automatically annotate proteins with high accuracy:
- UniRule (Manually curated rules)
- SAAS (System generated rules)

Supporting data
Select one of the options below to target your search:
- Literature citations
- Taxonomy
- Keywords
- Subcellular locations
- Cross-referenced databases
- Human diseases
Automatic annotation in UniProtKB

Unreviewed

select

Unreviewed

annotate

update

Expert curation

update

test against

Automatic annotation pipelines

integrate

Reviewed

expert rules

automatic rules
The Statistical Automatic Annotation System

• SAAS is a fully automatic annotation system.
• It is a machine learning algorithm that learns on expertly annotated entries in UniProtKB/Swiss-Prot entries to generate models allowing to automatically annotate TrEMBL.

![Diagram showing the process]

• The learning step in SAAS is performed by a decision tree (C4.5) classifier.
• Learning attributes: InterPro signatures, taxonomy, sequence length.
• The learning result is a tree from which annotation rules can be extracted. Then, these rules are applied on UniProtKB/TrEMBL entries.
SAAS Decision tree learning explained

<table>
<thead>
<tr>
<th>SwissProt</th>
<th>IPR005478</th>
<th>Eukaryota</th>
<th>PF02780</th>
<th>Transketolase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prot1</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Prot2</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Prot3</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Prot4</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Prot5</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Decision Tree Diagram:**

1. **IPR005478?**
   - Yes: Prot1, Prot4, Prot2, Prot5, Prot3
   - No: Prot2, Prot3

2. **PF02780?**
   - Yes: Prot1, Prot5
   - No: Prot4

**Transketolase**

- Yes
- No
SAAS: SAAS00001305

You are viewing a SAAS fully automatic annotation rule. When conditions on the left side are met, the relevant annotations on the right are applied to protein entries. Click here to search UniProt for all the entries that have been annotated by this rule.

If a protein meets these Conditions...

- TAXON = Bacteria
- INTERPRO_ID = IPR018357
- INTERPRO_ID = IPR029098
- SCOP_SUPERFAMILY_ID = SSF51161
- TIGRFAM_ID = TIGR01852

... then these Annotations are applied

Function
Involved in the biosynthesis of lipid A, a phosphorylated glycolipid that anchors the lipopolysaccharide to the outer membrane of the cell.
Sharing UniProt rules with the community
Project aims

• Sharing the UniProt annotation rules
  • To open the UniProt systems to the community
  • To facilitate the exchange of rules between research groups
• Providing a way to apply the rules on your data
  • Standalone tool
  • Service
• Open to contribution
  • We need your feedback
  • Open-source code, extensible model, …
Sharing UniProt rules with the community

Two components:

1. Exchange format:
   - Common for UniRule, SAAS (, and future) rules
   - Standard: well supported / understood
   - Exposing all the logic to interpret the rules

2. Engine:
   - Able to apply the rules on a set of data
   - Able to retrieve the inferred annotations
Executing the UniProt rules on your data
Register your interest

• Would you like to try out the UniProt rules to annotate your own data?
• Are you interested in integrating the UniProt rules (UniRule and SAAS) in your annotation pipeline?

Register your interest on the UniProt blog:

http://insideuniprot.blogspot.co.uk/2018/03/would-you-like-to-annotate-function.html

Or contact us directly at: help@uniprot.org
Accessing UniProt Data

• UniProt Releases every 4 weeks and is freely available

• Online training: http://www.ebi.ac.uk/training/online/course/uniprot-exploring-protein-sequence-and-functional

• Online training: https://www.ebi.ac.uk/training/online/course/interpro-functional-and-structural-analysis-protei

• Contact us: help@uniprot.org