Welcome - webinar instructions

• The webinar will start soon

• GoToTraining works best in **Chrome** or on Linux, **Firefox**

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

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

• Please complete the **feedback survey** which will launch at the end of the webinar

• The webinar will be recorded and added to Train online
Annotate your proteins with the UniProt Functional Annotation System UniFIRE

Hermann Zellner
hzellner@ebi.ac.uk
https://gitlab.ebi.ac.uk/uniprot-public/unifire
Predicting the function of protein sequences

• Growing number of protein sequences which function is unknown.
• Many methods inferring the function from databases of proteins whose functions are known, based on:
  • Sequence similarity
  • Protein families & domains
  • Taxonomy
  • Genomic context
  • ...

Combining methods and criteria

Annotation Rules
Project aims

• Annotation rules are valuable knowledge that should be publicly shared
• Distributing 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
  https://gitlab.ebi.ac.uk/uniprot-public/unifire
• UniFIRE mailing list:
  https://listserver.ebi.ac.uk/mailman/listinfo/unifire
The UniProt functional annotation systems
The Two Sides of UniProtKB

Reviewed
(UniProtKB/Swiss-Prot)
• Expert-curated literature annotation
• Reference for quality control of rule systems

Unreviewed
(UniProtKB/TrEMBL)
• Automatically annotated by rule systems
The UniRule System

UniRule: UR000000241

Source ID: RJ090372

If a protein meets these conditions...''

Common conditions

- Matches Pfam signature PRO140
- Matches PROSITE signature PRO1104
- Taxon = Archaea, Eubacteria

Special conditions

- Taxon = Archaea
- Taxon = Eubacteria

... then these annotations are applied''

Protein names

- Recommended name: 60S ribosomal protein L13a
- Recommended name: 60S ribosomal protein L13

Sequence similarities

Belongs to the eukaryotic ribosomal protein 6L13 family,

Keywords

- Ribonucleoprotein
- Ribosomal protein

Sample of UniProt entry J3QSB4 showing the evidence tags for a UniRule.

UniRule statistics for UniProtKB/TreMBL (release 2017_07)
UniRule: more complex example

If a protein meets these conditions...

Common conditions

- Matches HAMAP signature MF_01657
- taxon = Bacteria
- fragment + the sequence is fragmented

Special conditions

- taxon = Enterobacteriales
- Subsequence at position 12 - 15 aligns to "[ST]-G-[NDAVS]-[IV]" in entry Q52060 (individually applies "NAD")
- Subsequence at position 132 - 132 aligns to "C" in entry Q52060 (individually applies "Acyl-thioester intermediate")
- Subsequence at position 163 - 171 aligns to "[SG]-x-G-[PI]-[GA]-[TS]-x-x-N" in entry Q52060 (individually applies "NAD")
- Subsequence at position 290 - 290 aligns to "N" in entry Q52060 (individually applies "NAD")

... then these annotations are applied

- Recommended name:
  - Acetaldehyde dehydrogenase (EC:1.2.1.10)
  - Alternative name(s):
    - Acetaldehyde dehydrogenase [acetylating]

- Gene name:
  - Name: mhpF

- Catalytic activity:
  - Acetaldehyde + CoA + NAD⁺ → acetyl-CoA + NADH.

- Subunit structure:
  - Interacts with MhpE.

- Pathway:
  - 3-phenylpropionate degradation
    - This protein is involved in the pathway 3-phenylpropionate degradation, which is part of Aromatic compound metabolism.
    - View all proteins of this organism that are known to be involved in the pathway 3-phenylpropionate degradation and in Aromatic compound metabolism.

- Sequence similarities:
  - Belongs to the acetaldehyde dehydrogenase family.

- Function:
  - Catalyzes the conversion of acetaldehyde to acetyl-CoA, using NAD⁺ and coenzyme A. Is the final enzyme in the meta-cleavage pathway for the degradation of aromatic compounds.

- Active site:
  - Acyl-thioester intermediate (to residues corresponding to position 132)

- Nucleotide binding:
  - NAD (to residues corresponding to positions 12 - 15)
  - NAD (to residues corresponding to positions 163 - 171)

http://www.uniprot.org/unirule/UR000083169
The Statistical Automatic Annotation System

- Generates automatic rules for functional annotation from expertly annotated entries in UniProtKB/Swiss-Prot using the C4.5 decision tree algorithm
- Algorithm finds the most concise rule for an annotation based on the properties of sequence length, InterPro group membership and taxonomy
UniProt rules: conditions & annotations

**Conditions**
- InterPro Signatures:
  - (Super)families
  - Domains
  - Repeat
  - Site
- Organism:
  - Taxonomy
  - Scientific name
- Gene location (organelle)
- Sequence:
  - Length
  - Is fragmented
  - Patterns

**Annotations**
- Protein name
- Function
- Gene Ontology terms
- Catalytic activity
- Pathway membership
- Subcellular location
- Sequence positional features
  - positions of post-translational modifications
  - active sites
  - …
Sharing UniProt rules with the community
Sharing UniProt rules with the community

Two components:

1. Exchange format:
   - Common for UniRule, SAAS, (ARBA, ...) 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
The UniProt Rule Markup Language (URML)

Features:

- Based on XML
- Rules as the basic logical unit: (conditions → actions)
- Linked to a fact model
- Explicit condition querying using fact model class/attributes
- Annotations in the form of tuples: (type, value)

General structure:

```xml
name="org.uniprot.unirule" version="2018_02">
  <rule id="UR000101276_main" group="UR000101276">...</rule>
  <rule id="UR000101276_case_1" extends="UR000101276_main" group="UR000101276">...</rule>
  <rule id="UR000101276_case_2" extends="UR000101276_main" group="UR000101276">...</rule>
  <rule id="UR000101276_pfs_1" extends="UR000101276_main" group="UR000101276">...</rule>
  <rule id="UR000101276_pfs_1_pf_1" extends="UR000101276_pfs_1" group="UR000101276">...</rule>
</rules>
```

One UniRule (UR000101276) represented with 5 rules grouped with the same UniRule ID.
URML: The Fact Model

Positional feature support
<conditions>
  <AND>
    <condition on="fact:Protein" bind="protein">
      <filter on="sequence.isFragment" negative="true"/>
    </condition>
    <condition on="fact:Organism" bind="organism" of="protein">
      <filter on="lineage.ids">
        <contains operator="any">
          <value description="Bacteria">2</value>
        </contains>
      </filter>
    </condition>
    <condition on="fact:ProteinSignature" with="protein">
      <filter on="signature">
        <field attribute="type">HAMAP</field>
        <field attribute="value">MF_01224_B</field>
      </filter>
    </condition>
  </AND>
  <AND>
    <condition on="fact:Protein" bind="protein">
      <filter on="sequence.isFragment" negative="true"/>
    </condition>
    <condition on="fact:Organism" bind="organism" of="protein">
      <filter on="lineage.ids">
        <contains operator="any">
          <value description="Archaea">2157</value>
        </contains>
      </filter>
    </condition>
    <condition on="fact:ProteinSignature" with="protein">
      <filter on="signature">
        <field attribute="type">HAMAP</field>
        <field attribute="value">MF_01224_A</field>
      </filter>
    </condition>
  </AND>
</conditions>
<conditions>
  <AND>
    <condition on="fact:Protein" bind="protein">
      <filter on="sequence.isFragment" negative="true"/>
    </condition>
  </AND>
  <AND>
    <condition on="fact:Organism" bind="organism" of="protein">
      <filter on="lineage.ids">
        <contains operator="any">
          <value description="Bacteria">2</value>
        </contains>
      </filter>
    </condition>
    <condition on="fact:ProteinSignature" with="protein">
      <filter on="signature">
        <field attribute="type">HAMP</field>
        <field attribute="value">MF_01224_B</field>
      </filter>
    </condition>
  </AND>
  <AND>
    <condition on="fact:Protein" bind="protein">
      <filter on="sequence.isFragment" negative="true"/>
    </condition>
    <condition on="fact:Organism" bind="organism" of="protein">
      <filter on="lineage.ids">
        <contains operator="any">
          <value description="Archaea">2157</value>
        </contains>
      </filter>
    </condition>
    <condition on="fact:ProteinSignature" with="protein">
      <filter on="signature">
        <field attribute="type">HAMP</field>
        <field attribute="value">MF_01224_A</field>
      </filter>
    </condition>
  </AND>
</conditions>
URML: Condition Space

If there is a fact Protein (that we call “protein”) having its attribute “sequence” having its attribute “isFragment” == false

And the Organism of that protein has its attribute “lineage” having its attribute “ids” that contains 2157.

And there is a ProteinSignature with protein having the attribute “signature” having its attributes: “type”=HAMAP and “name”=MF_01224_A

Then …
URML: Action Space

<actions>
  <action type="create" with="evidence:'UR000101276' protein">
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">comment.catalytic_activity</field>
      <field attribute="value">Reaction=(8S)-3',8-cyclo-7,8-dihydroguanosine 5'-triphosphate = cyclic pyranopterin phosphate + diphosphate; Xref=Rhea:RHEA:49580, ChEBI:CHEBI:33019, ChEBI:CHEBI:59648, ChEBI:CHEBI:131766; EC=4.6.1.17</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">comment.function</field>
      <field attribute="value">Catalyzes the conversion of (8S)-3',8-cyclo-7,8-dihydroguanosine 5'-triphosphate to cyclic pyranopterin monophosphate (cPMP)</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">comment.pathway</field>
      <field attribute="value">Cofactor biosynthesis; molybdopterin biosynthesis</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">comment.similarity</field>
      <field attribute="value">Belongs to the MoaC family</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">comment.subunit</field>
      <field attribute="value">Homohexamer; trimer of dimers</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">keyword</field>
      <field attribute="value">Lyase</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">keyword</field>
      <field attribute="value">Molybdenum cofactor biosynthesis</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">gene.name.primary</field>
      <field attribute="value">moaC</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">xref.GO</field>
      <field attribute="value">GO:0066777</field>
    </fact>
    <fact type="fact:ProteinAnnotation">
      <field attribute="type">xref.GO</field>
      <field attribute="value">60:0061799</field>
    </fact>
  </action>
</actions>
Create ProteinAnnotation facts, with type and value.
Every ProteinAnnotation is bound to its protein and an evidence (rule id)

Create ProteinAnnotation facts, with type and value.
UniFIRE: The UniProt Functional annotation Inference Rule Engine
Input / Output

Input Sequences (MultiFasta)

InterProScan

InterProScan Output (XML)

InterProScan

Rule File (URML)

Template alignments (FACT)

UniFIRE

Annotations (XML / TSV)

Fasta header containing all required metadata.

Alternative input: using the fact model
UniFIRE Gitlab Repository

https://gitlab.ebi.ac.uk/uniprot-public/unifire

---

**UniFIRE**

Project ID: 1200

Apache License 2.0  -  4 Commits  -  1 Branch  -  0 Tags  -  15.5 MB Files

**Details**

- **master**  
- **unifire**

Use `$(uniprot-dependencyversion) property for unirule.public dependency`  
Hermann Zellner authored 1 week ago

**README**  
**CHANGELOG**

<table>
<thead>
<tr>
<th>Name</th>
<th>Last commit</th>
<th>Last update</th>
</tr>
</thead>
<tbody>
<tr>
<td>core</td>
<td>Initial commit of UniFIRE codebase</td>
<td>1 week ago</td>
</tr>
<tr>
<td>distribution</td>
<td>Initial commit of UniFIRE codebase</td>
<td>1 week ago</td>
</tr>
<tr>
<td>engine</td>
<td>Initial commit of UniFIRE codebase</td>
<td>1 week ago</td>
</tr>
<tr>
<td>io</td>
<td>Initial commit of UniFIRE codebase</td>
<td>1 week ago</td>
</tr>
<tr>
<td>misc</td>
<td>Initial commit of UniFIRE codebase</td>
<td>1 week ago</td>
</tr>
</tbody>
</table>
UniFIRE (The UniProt Functional annotation Inference Rule Engine) is a software to execute rules in the UniProt Rule Markup Language (URML) format. This software can predict functional annotations for protein sequences by using the UniProt annotation rules (UniRule and SAAS).

Please have a look at the slides of the introducing presentation for further information about this project.

Examples how to use UniFIRE software, can be found in the README.md.

Information about new features, updates and changes to the code or repository can be found in the News page.


<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Last Modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>File: saas-urml-2019_06.xml</td>
<td>36097 KB</td>
<td>14/06/2019 10:36:00 BST</td>
</tr>
<tr>
<td>saas-urml-latest.xml</td>
<td>1571 KB</td>
<td>14/06/2019 10:36:00 BST</td>
</tr>
<tr>
<td>File: unirule-templates-2019_06.xml</td>
<td>36481 KB</td>
<td>14/06/2019 10:36:00 BST</td>
</tr>
<tr>
<td>unirule-templates-latest.xml</td>
<td></td>
<td></td>
</tr>
<tr>
<td>File: unirule-urml-2019_06.xml</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unirule-urml-latest.xml</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Demo of UniFIRE
Demo – Input Sequence

Protein sequence taken from Richter et al., J Mol Biol., 2010

Overview

<table>
<thead>
<tr>
<th>Entry</th>
<th>Protein names</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGAAKINL68</td>
<td>Imidazole glycerol phosphate synthase subunit HisF (Desulfurobacterium indicum)</td>
</tr>
<tr>
<td>EBT344</td>
<td>Imidazole glycerol phosphate synthase subunit HisF (Thermus thermophilus)</td>
</tr>
<tr>
<td>AAMAK3</td>
<td>Imidazole glycerol phosphate synthase subunit HisF (Isotoma isota)</td>
</tr>
<tr>
<td>B11556</td>
<td>Imidazole glycerol phosphate synthase subunit HisF (Desulfurococcus aurantiaca)</td>
</tr>
</tbody>
</table>

Alignments

76% identity not enough to infer annotations by similarity
Demo – Running InterProScan

https://www.ebi.ac.uk/interpro/search/sequence-search

interproscan-output.xml
Demo – Download and build UniFIRE

Clone the UniFIRE gitlab repository:

```bash
$ git clone https://gitlab.ebi.ac.uk/uniprot-public/unifire
Cloning into 'unifire'...
warning: redirecting to https://gitlab.ebi.ac.uk/uniprot-public/unifire.git/
remote: Enumerating objects: 300, done.
remote: Counting objects: 100% (300/300), done.
remote: Compressing objects: 100% (157/157), done.
remote: Total 300 (delta 62), reused 297 (delta 59)
Receiving objects: 100% (300/300), 15.43 MiB | 4.79 MiB/s, done.
Resolving deltas: 100% (62/62), done.
```

Use build.sh script to download dependencies and build UniFIRE with maven build tool. It also downloads the latest UniRule and SAAS rule from EBI FTP files into the folder unifire/samples

```bash
$ cd unifire
$ ./build.sh
Building UniFIRE and downloading dependencies...
```
Demo – Applying UniRule rules with UniFIRE

```
$ ./distribution/bin/unifire.sh -r samples/unirule-urml-latest.xml -t samples/unirule-templates-latest.xml -i mystery_lineage.xml -o mystery_lineage.out
```

17:21:55.690 INFO u.a.e.u.u.e.d.c.DroolsCompiler - Compiling rule base org.uniprot.unirule-2019_06
17:23:18.060 INFO u.a.e.u.u.c.x.r.URMLFactReader - Reading input facts
17:23:18.337 INFO u.a.e.u.u.UniFireRunner - Fire all rules on 1 protein(s)
17:23:19.733 INFO u.a.e.u.u.UniFireRunner - Write 25 prediction(s)

```xml
<fact xsi:type="ProteinAnnotation">
  <protein>LUCAC1 0001</protein>
  <evidence>UR000100401</evidence>
  <type>feature.ACT_SITE</type>
  <positionStart>11</positionStart>
  <positionEnd>11</positionEnd>
</fact>

<fact xsi:type="ProteinAnnotation">
  <protein>LUCAC1 0001</protein>
  <evidence>UR000100401</evidence>
  <type>keyword</type>
  <value>Lyase</value>
</fact>

<fact xsi:type="ProteinAnnotation">
  <protein>LUCAC1 0001</protein>
  <evidence>UR000100401</evidence>
  <type>protein.recommendedName.ecNumber</type>
  <value>4.3.2.18</value>
</fact>

<fact xsi:type="ProteinAnnotation">
  <protein>LUCAC1 0001</protein>
  <evidence>UR000100401</evidence>
  <type>comment.subunit</type>
  <value>Heterodimer of HisH and HisF</value>
</fact>

<table>
<thead>
<tr>
<th>Evidence</th>
<th>ProteinId</th>
<th>AnnotationType</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>comment.catalytic activity</td>
<td>Reaction: S^5-(phospho-1-deoxy-D-ribose-1-ulose-5-phosphate)</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>comment.function</td>
<td>IGPS catalyzes the conversion of PRPP and glutamine to GIP and glutamate. The HisF subunit catalyzes the cyclization activity that produces GIP and ACIP from IGPS using the ammonia provided by the HisF subunit.</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>comment.pathway</td>
<td>Belongs to the HisA/HisF family</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>comment.similarity</td>
<td>Belongs to the HisA/HisF family</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>comment.similarity</td>
<td>CpG</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>comment.subcellular location</td>
<td>Cytoplasm</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>comment.subunit</td>
<td>Heterodimer of HisH and HisF</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>gene_product</td>
<td>HisF</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
<tr>
<td>UR000100401</td>
<td>LUCAC1 0001</td>
<td>keyword</td>
<td>Amino-acid biosynthesis</td>
</tr>
</tbody>
</table>
```
Demo – Let’s look at the triggered rule

If a protein meets these conditions...

Common conditions

- Matches HAMAP signature MF_01013
- Domain = Bacteria, Archaea
- Fragment: the sequence is fragmented

... then these annotations are applied...

Protein name

- Recommended name: Imidazole glycerophosphate synthase subunit HisF (EC:4.3.2.10)
- Alternative name(s):
  - IGP synthase cyclase subunit
  - IGP synthase subunit HisF
  - ImGP synthase subunit HisF
- Short name: IGP subunit HisF

Gene name

- Name: hisF

Subunit structure

- Heterodimer of HisA and HisF.

Catalytic activity

- (5-phospho-1-deoxy-0-ribulos-1-ylimino)methylamine-1-(5-phospho-β-D-ribofuranosyl)imidazole-4-carboxamide + L-glutamine + S-amino-1-(5-phospho-β-D-ribofuranosyl)imidazole-4-carboxamide + D-erythro-1-(imidazol-4-yl)glycerol 3-phosphate + H^+ + L-glutamate
- EC:4.3.2.10
- Source: Rhea

Sequence similarities

- Belongs to the HisA/HisF family.

Subcellular location

- Cytoplasm

Function

- IGP catalyzes the conversion of PRPR and glutamine to IGD, AISCAR and glutamate. The HisF subunit catalyzes the cyclization activity that produces IGP and AISCAR from PRPR using the ammonia provided by the HisF subunit.

Pathway


Summary

We want to exchange the knowledge about functional annotation of proteins

- We have designed:
  - A format to exchange annotation rules
  - An engine to apply them

- We would need your feedback and ideas:
  - On the rule format
  - On the input / output data
  - On the engine

UniFIRE mailing list: https://listserver.ebi.ac.uk/mailman/listinfo/unifire
Thank you for your attention!
Any questions?
Upcoming webinars

See the full list of upcoming webinars at https://www.ebi.ac.uk/training/webinars

Don’t forget!

Please fill in the survey that launches after the webinar – thanks!