Towards automating the ontological representation of proteins in the Protein Ontology

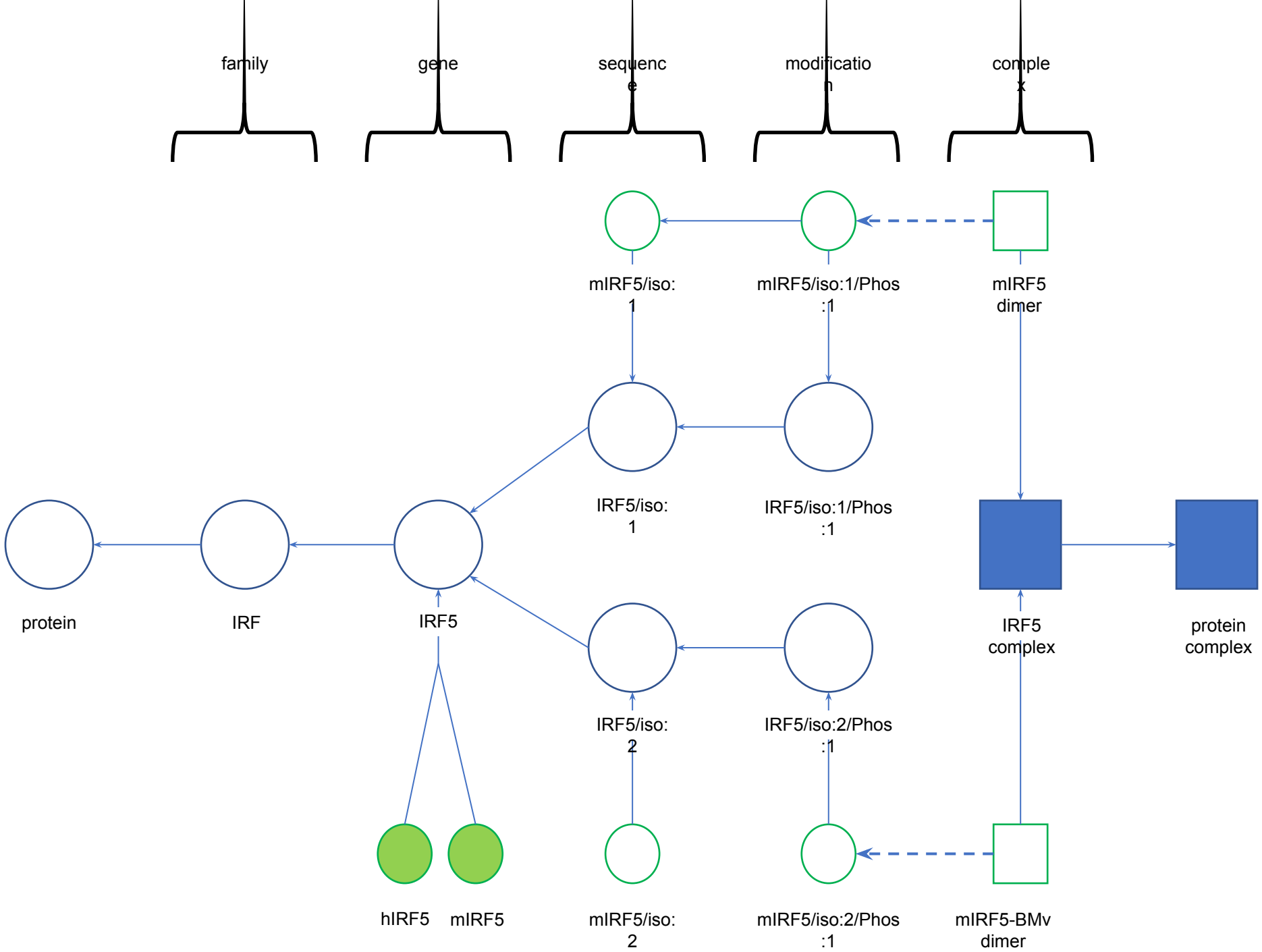
Workshop on Synergizing Biomedical Ontologies

July 14-16, 2021

Darren A. Natale

Smad 2	“canonical” isoform	•Cytoplasmic	SMAD2_HUMAN UniProtKB:Q15796-1
Smad 2	alternatively spliced short form	•Forms complex •Nuclear •Txn upregulation	SMAD2_HUMAN UniProtKB:Q15796-2
Smad 2 x	point mutation (causative agent: large intestine carcinoma)	•Forms complex •Nuclear •Txn upregulation++	SMAD2_HUMAN VAR_011375
Smad 2 P P	TGF-β receptor phosphorylated	•Forms complex •Cytoplasmic •No Txn upregulation	
P P P P P Smad 2 P P	TGFBR+ERK1 phosphorylated	•Cytoplasmic	
P Smad 2 P P	TGFBR+CAMK2 phosphorylated	•Nuclear •Txn upregulation	
Smad 2 P P	TGF-β receptor phosphorylated short form	•Doesn’t form complex •Cytoplasmic •No Txn upregulation	

UniProtKB:Q15796


family
gene
sequenc
e
modificatio
n
comple
x
mIRF5/iso:
1
mIRF5/iso:1/Phos
:1
mIRF5
dimer
IRF5/iso:
1
IRF5/iso:1/Phos
:1
protein
IRF
IRF5
IRF5
complex
protein
complex
IRF5/iso:
2
IRF5/iso:2/Phos
:1
hIRF5
mIRF5
mIRF5/iso:
2
mIRF5/iso:2/Phos
:1
mIRF5-BMv
dimer

What, precisely, is an ortho-proteoform?

same gene

```
Q13568-1 IRF5_HUMAN 360 PIQREVKTKLFSLEHFLNELILFQKGQTNTPPPFEIFFCFGEEWPDRKPREKKLITVQVV 419
P56477-1 IRF5_MOUSE 359 PIQREVKTKLFSLEQFLNELILFQKGQTNTPPPFEIFFCFGEEWPDVKPREKKLITVQVV 418
                        ***************:****************************** *************

Q13568-1 IRF5_HUMAN 420 PVAARLLLEMFSGELSWSADSIRLQISNPDLKDRMVEQFKELHHIWQSQQRLQPVAQAPP 479
P56477-1 IRF5_MOUSE 419 PVAARLLLEMFSGELSWSADSIRLQISNPDLKDHMVEQFKELHHLWQSQQQLQPMVQAPP 478
                        ********************************:**********:*****:***:.****

Q13568-1 IRF5_HUMAN 480 GAGLGVGQGPWPMHPAGMQ                                          498
P56477-1 IRF5_MOUSE 479 VAGLDASQGPWPMHPVGMQ                                          497
                         ***...********.***
```

same modifications

same isoform

Automating the creation of UniProtKB-based terms

Eligible:

- Proteins from human, mouse, rat, chicken, zebrafish, worm, fruit fly, slime mold, budding and fission yeasts, *Arabidopsis*, *E. coli*
- Reviewed (UniProtKB/Swiss-Prot) entries

Key Information:

- Protein name & synonyms
- Organism
- Gene name & synonyms

Other leverageable information:

- Type of sequence (isoform, variant)
- Sequence processing & PTMs

Technical policies:

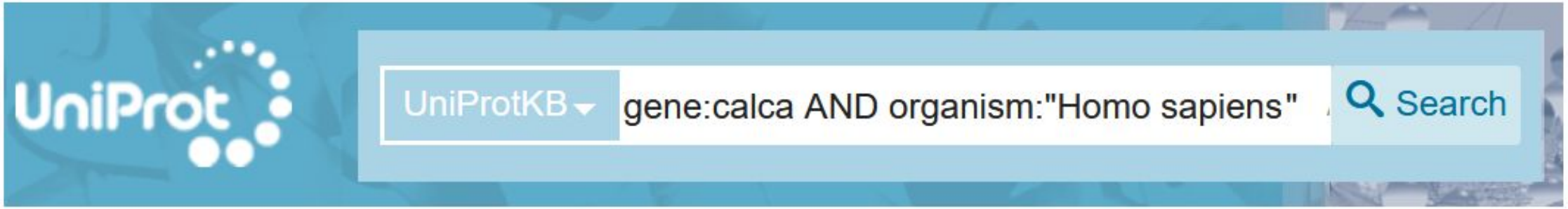
- Well-described format
- Obsolescence

What, precisely, is a UniProtKB entry?

Whenever possible, all the protein products encoded by one gene in a given species are described in a single UniProtKB/Swiss-Prot entry, including isoforms generated by alternative splicing, alternative promoter usage, and alternative translation initiation (*). However, some alternative splicing isoforms derived from the same gene share only a few exons, if any at all, the same for some 'trans-splicing' events. In these cases, the divergence is obviously too important to merge all protein sequences into a single entry and the isoforms have to be described in separate 'external' entries.

(*) Important remark: Due to the increase of sequence data coming from large-scale sequencing projects, UniProtKB/TrEMBL may contain additional predicted sequences encoded by genes which are described in a UniProtKB/Swiss-Prot entry.

Example of a “split” UniProtKB entry



P01258	CALC_HUMAN		**Calcitonin**	**CALCA**	CALC1	Homo sapiens	141
P06881	CALCA_HUMAN		**Calcitonin gene-related peptide 1**	**CALCA**	CALC1	Homo sapiens	128

```
P01258 CALC_HUMAN     1 MGFQKFSPFLALSILVLLQAGSLHAAPFRSALESSPADPATLSEDEARLLLAALVQDYVQ   60
P06881 CALCA_HUMAN    1 MGFQKFSPFLALSILVLLQAGSLHAAPFRSALESSPADPATLSEDEARLLLAALVQDYVQ   60
                        ************************************************************

P01258 CALC_HUMAN    61 MKASELEQEQEREGSSLDSPRSKRCGNLSTCMLGTYTQDFNK-----FHTFPQTAIGVGA  115
P06881 CALCA_HUMAN   61 MKASELEQEQEREGSRIIAQKR-AC-DTATCVTHRLAGLLSRSGGVVKNNFVPTNVGSKA  118
                        *************** : : :   * : :**:    :  :.:      :.*  * :*  *

P01258 CALC_HUMAN   116 PGKKR-DMSSDLERDHRPHVSMPQNAN                                   141
P06881 CALCA_HUMAN  119 FGRRRRDLQA-----------------                                   128
                         *::* *:.:
```

How PRO handles a “split” UniProtKB entry

PR:000027222	*CALCA gene translation product*
PR:000030026	*CALCA gene translation product (human)*
PR:P06881-1	*calcitonin gene-related peptide 1 (human)*
PR:P01258	*calcitonin isoforms 1/2 (human)*
PR:P01258-1	*calcitonin isoform 1 (human)*
PR:P01258-2	*calcitonin isoform h2 (human)*
PR:000027223	*CALCA gene translation product (mouse)*
PR:000036869	*CALCA gene translation product (rat)*

1) Create a “catch-all” term for each organism, the parent for all protein products from that gene in that organism
2) As these are actually isoforms, use the isoform designation
3) If a split entry has multiple isoforms, treat the “main” entry as a defined class

PRO treatment of TrEMBL entries

Might represent:

1) Existing isoform (already in a Swiss-Prot entry)
2) New isoform (for a protein already in Swiss-Prot)
3) Protein with no Swiss-Prot entry for the given gene
    a) Sole entry for that gene
    b) One of several possible
        i. Canonical
        ii. Alternative splice form
4) Natural variant of a sequence

PRO dynamic term generator

https://proconsortium.org/

Retrieve a PRO entry (enter a PRO ID):

Example: PR:000025934 (sample output)

Enter UniProtKB accession

UniProt APIs

Protein Ontology report - protein lin-7 homolog A isoform M0R7K1 (rat)

M0R7K1 - http://purl.obolibrary.org/obo/PR_M0R7K1

This term was generated dynamically and will not appear in any downloadable version of PRO unless such is requested. Please note that a github login is required.

Show OBO stanza

Ontology Information	
PRO ID	PR:M0R7K1
PRO name	protein lin-7 homolog A isoform M0R7K1 (rat)

How it works: the queries

1. On-the-fly retrieval of the UniProtKB information file for the input accession.
2. Check to see if the accession for the returned entry is the same as input.
3. A query through UniProt's UniParc database to see which other accessions the requested one might be known by.
4. A gene-based query through UniProtKB to determine if there are multiple possible entries for that gene

```
Q9H633-2    RPP21_HUMAN        1 MAGPVKDREAFQRLNFLYQVSLRQGPHGDGARRPRVTAPLPQAAHCVLAQDPENQALARF  60
A0A0G2JJ52  A0A0G2JJ52_HUMAN   1 MAGPVKDREAFQRLNFLYQVSLRQGPHGDGARRPRVTAPLPQAAHCVLAQDPENQALARF  60
                                 ************************************************************

Q9H633-2    RPP21_HUMAN       61 YCYTERTIAKRLVLRRDPSVKRTLCRGCSSLLVPGLTCTQRQRRCRGQRWTVQTCLTCQR 120
A0A0G2JJ52  A0A0G2JJ52_HUMAN  61 YCYTERTIAKRLVLRRDPSVKRTLCRGCSSLLVPGLTCTQRQRRCRGQRWTVQTCLTCQR 120
                                 ************************************************************

Q9H633-2    RPP21_HUMAN      121 SQRFLNDPGHLLWGDRPEAQLGSQADSKPLQPLPNTAHSISDRLPEEKMQTQGSSNQ    177
A0A0G2JJ52  A0A0G2JJ52_HUMAN 121 SQRFLNDPGHLLWGDRPEAQLGSQADSKPLQPLPNTAHSISDRLPEEKMQTKGSSNQ    177
                                 ***************************************************:*****
```

The next challenge: dealing with modified/variant proteins

Prerequisites:

1) Known UniProtKB entry
2) Known range of sequence
3) Known positions and type of modification

Issues:

1) Disparities in sequence numbering or isoforms used
2) Disparities in how modifications are indicated

3) Identifying existing terms

If isoform 1 of UniProtKB:P12345 is canonical, **the following are all equivalent**:

- UniProtKB:P12345, Ser-77, MOD:00696
- UniProtKB:P12345, Ser-77, MOD:00046
- UniProtKB:P12345, Ser-77, CHEBI:45522
- UniProtKB:P12345-1, Ser-77, MOD:00696
- UniProtKB:P12345-1, Ser-77, MOD:00046
- UniProtKB:P12345-1, Ser-77, CHEBI:45522

PRO/UniProt/Reactome: full canonical
Alzforum: ‘favorite’ isoform
Histome: removed InitMet

PRO: PSI-MOD, GNOme
UniProt: internal controlled vocabulary
Reactome/TDR: PSI-MOD, CHEBI

MOD:00696 = phosphorylated residue
MOD:00046 = phosphorylated serine
CHEBI:45522 = phosphorylated serine