RACE-PRO: Rapid Annotation interfaCE for Protein Ontology
Quick guide

What is RACE-PRO for?

- Obtain a PRO ID for the protein objects of interest. For example, if you are modeling pathways you would like to be able to represent and refer to all the different forms of a given protein in order to describe the various reactions in the pathway. Each of these forms is a protein object in PRO and will have a distinct ID (e.g. PRO ID for smad2 isoform 1 and a PRO ID for smad2 isoform 1 phosphorylated in a given residue). By creating a RACE-PRO entry, you are not only requesting a PRO ID for a specific protein form, but also assisting in the curation process.
- Define a protein object (based on literature and/or experimental data). The literature or the experiment may describe a set of proteins and protein forms (splice variants that undergo some modification). For example, in the processing of a protein by cleavage, usually the papers refer to a precursor and a mature form of such protein. Each of these can be defined using a sequence and defining the regions (subsequence) corresponding to them. The paper could also describe some modification that applies to any of these forms, and those are also part of the definition of the protein object.
- Add annotation to that protein object. Currently, in most databases the annotation is added to the canonical protein. There is hardly distinction about isoforms and/or modified forms. In RACE-PRO, the annotation is added to the most appropriate protein form, therefore if the paper shows that only a phosphorylated form of isoform 2 of protein x is localized to the nucleus, then this annotation is added only to the RACE-PRO entry for the phosphorylated form of isoform 2, and not to others. Only experimental information is added. Another important consideration is that the RACE-PRO entry goes by sequence and species. So any information that is entered in the entry has to be pertinent to both.

How does it work?

- Input your personal information (only for internal use)
- Complete form with sequence information and annotation
- Submit when ready (otherwise you can save for later)
- PRO curation team will take the data, revise it, and create the corresponding PRO node in the ontology with the corresponding source attribution
- Use will be informed through email about the new PRO IDs and when they will be public

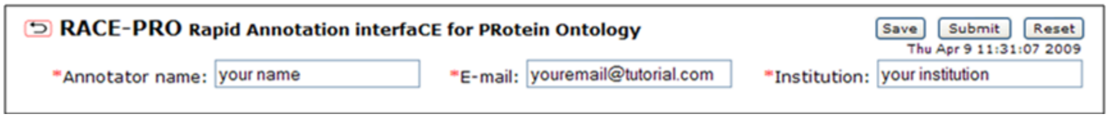
Protein A *n* protein forms=*n* protein objects=*n* RACE-PRO entries

It is important to note that in RACE-PRO a protein object refers to a distinct protein form. So if there are two different forms of a given protein (e.g. unmodified smad2 *vs.* phosphorylated smad2), two RACE-PRO records are created. Similarly, if a paper describes a protein for which more than one modification co-exists (e.g. the active form of smad2 human is the protein that is phosphorylated both in Ser-465 and Ser-467) then you would create 1 RACE-PRO entry (1 protein object), however the paper may describe modifications that occur under different events (e.g. the active form of smad2 described above vs. the inactive form phosphorylated in Ser-240, Ser-465and Ser-467), and therefore two protein forms (active and inactive) and two RACE-PRO entries.

It is recommended that before you start creating your RACE-PRO record that you search in PRO if there is any existing node that you can referenced to (in the comment section see 5.).

Creating a RACE-PRO entry

Fill your personal information. This information will not be distributed to any third party, it is only for saving your data and for communication purposes.



RACE-PRO Rapid Annotation interfaCE for PRotein Ontology

Thu Apr 9 11:31:07 2009

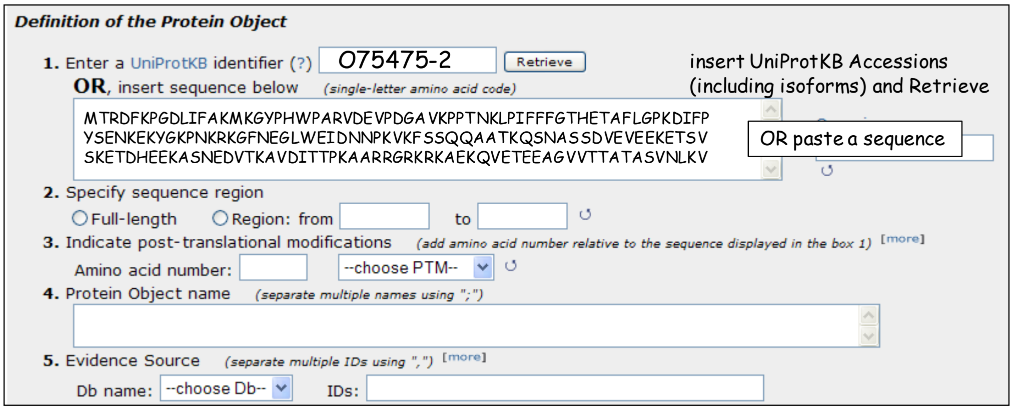
*Annotator name: your name

*E-mail: youremail@tutorial.com

*Institution: your institution

First Block: Definition of the protein object

1) Define the protein object. This block allows you to enter all the information about a protein form along with the source of evidence. Starting by retrieving or pasting a sequence, then defining the sequence length (is it the full length protein or is it a fragment or cleaved product that I want to define?), and post-translational modifications if present, and finally provide a name for the object (active form, mature protein, isoform delta1, etc) and source of evidence (most likely a Pubmed ID or PMID).



Definition of the Protein Object

1. Enter a UniProtKB identifier (?) O75475-2 Retrieve

OR, insert sequence below *(single-letter amino acid code)*

```
MTRDFKPGDLIFAKMKGYPHWPARVDEVPDGAVKPPTNKLPIFFFGTHETAFLGPKDIFP
YSENKEKYGKPNKRKGFNEGLWEIDNNPKVKFSSQQAATKQSNASSDVEVEEKETSV
SKETDHEEKASNEDVTKAVDITTPKAARRGRKRKAEKQVETEEAGVVTTATASVNLKV
```

2. Specify sequence region

○ Full-length ○ Region: from to

3. Indicate post-translational modifications *(add amino acid number relative to the sequence displayed in the box 1)* [more]

Amino acid number: --choose PTM--

4. Protein Object name *(separate multiple names using ";")*

5. Evidence Source *(separate multiple IDs using ",")* [more]

Db name: --choose Db-- IDs:

insert UniProtKB Accessions (including isoforms) and Retrieve

OR paste a sequence

1. *Retrieve the sequence:* If you use a UniProtKB identifier and click "Retrieve", the sequence retrieved is formatted to show the residue numbers, and the organism box is automatically filled. You can use identifiers for isoforms as the example shown above (a UniProtKB accession followed by a dash and a number). If you happen to have an identifier from a different database (Genebank ac, etc) you can use the ID mapping or batch retrieval service from the PIR main menu (under searh/analysis) to obtain the UniProtKB accession. However, it is safer to paste the sequence in this case unless you know what sequence within the UniProtKB record your accession corresponds to. A UniProtKB/Swiss-Prot record may contain more than one sequence which comes from different sources and may correspond to different isoforms, genetic variants, etc. Please be aware of this issue. If you use the ID mapping service you will link to the canonical sequence, but that accession could correspond to an isoform or a variant described in the file. So check the UniProtKB record.

If you paste a sequence, you need to reformat so you can see the residue numbers. Use the circle arrow to do so. Also you will need to add the organism, if you don't know the exact name you can follow the link to NCBI taxonomy browser by clicking on the Organism title.

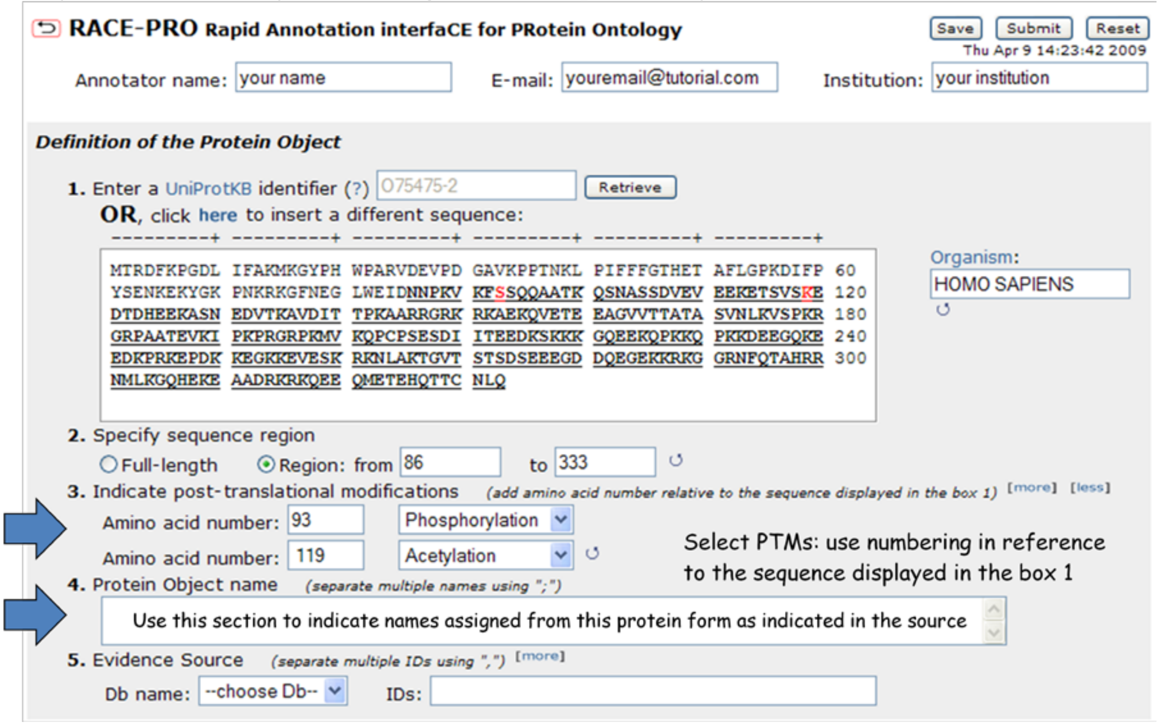
2.Protein region: After you have the sequence displayed in the box, you can select a subsequence in the cases where the protein form you are describing is a cleaved product or a fragment. After you do this, click on the circle arrow and your selected subsequence will be underlined. In the screenshot shown below, the protein object is a fragment of isoform 2 of O75475, which corresponds to the region starting at amino acid 86 and ending at amino acid 333.

3.Selecting the Modification: If you need to describe a modification (or modifications), enter the residue number and the type of modification. If the modification you need to enter is not in the list, use the "Other" option to add it. These terms will be later mapped to the corresponding PSI-MOD terms. For example, a serine phosphorylation, will be translated into MOD:00046 phospho-L-serine. If the modification site is unknown, please enter "?" in the residue number box.

Use the [more] to add another modification, and the [less] to remove one.

Be aware that the amino acid number will always refer to the sequence displayed in the sequence box. In the screenshot shown below, there are two modifications for this protein fragment, a phosphorylation and an acetylation the amino acid numbers correspond to the full sequence displayed in the sequence box. When clicking on the circle arrow, you will

see the residues highlighted. Please make sure that those are the expected residues. Note that if you select a residue for modification that is outside the region selected (for example, acetylation of Lysine 6) it will show a warning. Check that region and modification entered are correct.
4.*Protein object name*: Add names by which this object is known in the paper or source that you are using (separated by;)
5.*DB name*: add the database (DB) source of your annotation. Select one of the options in the list, if not present use the “Other” option and provide the name of the DB. In the ID box you can add many IDs for a given DB separated by comma.



RACE-PRO Rapid Annotation interfaCE for PRotein Ontology

Thu Apr 9 14:23:42 2009

Annotator name: your name E-mail: youremail@tutorial.com Institution: your institution

Definition of the Protein Object

1. Enter a UniProtKB identifier (?) O75475-2
OR, click here to insert a different sequence:

```
---------+ ---------+ ---------+ ---------+ ---------+ ---------+
MTRDFKPGDL IFAKMKGYPH WPARVDEVPD GAVKPPTNKL PIFFFGTHET AFLGPKDIFP 60
YSENKEKYGK PNKRKGFNEG LWEIDNNPKV KFSSQQAATK QSNASSDVEV EEKETSVSKE 120
DTDHEEKASN EDVTKAVDIT TPKAARRGRK RKAEKQVETE EAGVVTTATA SVNLKVSPKR 180
GRPAATEVKI PKPRGRPKMV KQPCPSESDI ITEEDKSKKK GQEEKQPKKQ PKKDEEGQKE 240
EDKPRKEPDK KEGKKEVESK RKNLAKTGVT STSDSEEEGD DQEGEKKRKG GRNFQTAHRR 300
NMLKGQHEKE AADRKRKQEE QMETEHQTTC NLQ
```

Organism: HOMO SAPIENS

2. Specify sequence region
○ Full-length ● Region: from 86 to 333

3. Indicate post-translational modifications *(add amino acid number relative to the sequence displayed in the box 1)*
Amino acid number: 93 Phosphorylation
Amino acid number: 119 Acetylation

Select PTMs: use numbering in reference to the sequence displayed in the box 1

4. Protein Object name *(separate multiple names using ";")*
Use this section to indicate names assigned from this protein form as indicated in the source

5. Evidence Source *(separate multiple IDs using ",")*
Db name: --choose Db-- IDs:

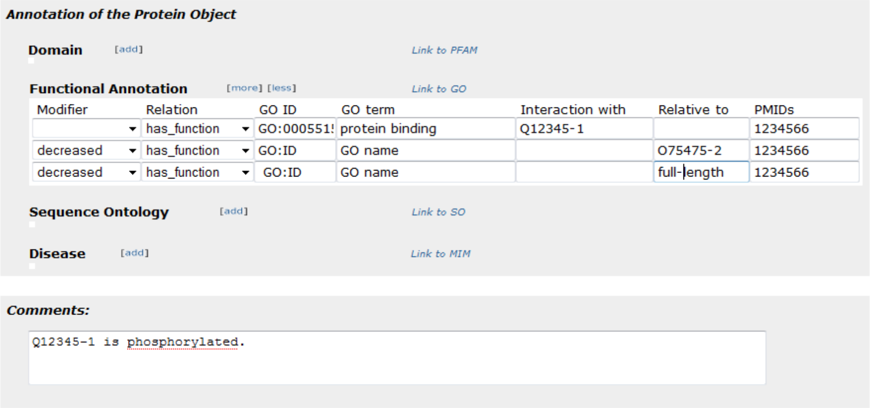
Second block: the annotation.
The annotations are separated by database/ontology: domain, functional terms, sequence and disease. You only add those with experimental data that means that in a given paper you look mainly at the Result section, especially the figures. Only annotation that is pertinent for the protein form (and species) described in block 1 should be added. For a tutorial for GO annotation please see (http://pir.georgetown.edu/~arighic/pro/tutorial/GO_Annotation_PRO.ppt).

All the information about the different columns in the table is in the PAF guidelines (ftp://ftp.pir.georgetown.edu/databases/ontology/pro_obo/PAF_guidelines.pdf). But below are some clarifications:

- Modifiers: used to modify a relation between a PRO term and another term. It includes the GO qualifiers NOT, contributes_to plus increased, decreased, and altered (to be used with the relative to column).
- Relation to the specific annotation. For some database/ontology there is a single relation possible and therefore it is already displayed
- Add ID for the specific database/ontology. If you need to search use the

"link to .." link. Future development: autofill of name.

- Interaction with column is used with the GO term protein binding to refer to the binding partner. Please add a UniProtKB Ac and/or PRO ID. If the binding partner happens to be modified please add this information in the comment (see first row and comment in the figure below).
- Relative to column is used only when a modifier such as increased, decreased or altered is used. You are expected to provide a reference to what entity the function is modified. Therefore, either provide a UniProtKB acc, the REF number (number assigned to your submitted entries), or the name. See two examples below: you could say that for the fragment that has been used as example above, the decreased function is in relation to the isoform 2 (using UniProtKB AC) or just enter full-length.



Annotation of the Protein Object

Domain [add] *Link to PFAM*

Functional Annotation [more] [less] *Link to GO*

Modifier	Relation	GO ID	GO term	Interaction with	Relative to	PMIDs
	has_function	GO:000551!	protein binding	Q12345-1		1234566
decreased	has_function	GO:ID	GO name		O75475-2	1234566
decreased	has_function	GO:ID	GO name		full-length	1234566

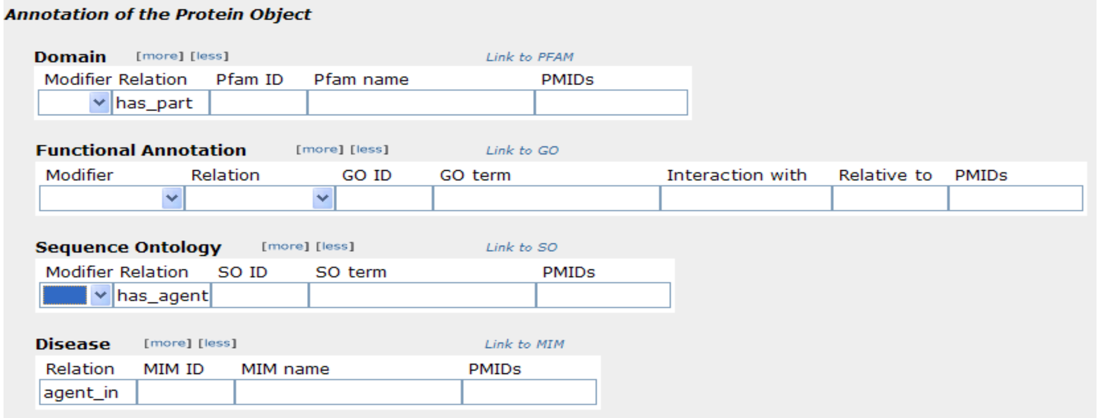
Sequence Ontology [add] *Link to SO*

Disease [add] *Link to MIM*

Comments:

Q12345-1 is phosphorylated.

Use the [more], [less] to add or remove an annotation line.



Annotation of the Protein Object

Domain [more] [less] *Link to PFAM*

Modifier	Relation	Pfam ID	Pfam name	PMIDs
	has_part			

Functional Annotation [more] [less] *Link to GO*

Modifier	Relation	GO ID	GO term	Interaction with	Relative to	PMIDs

Sequence Ontology [more] [less] *Link to SO*

Modifier	Relation	SO ID	SO term	PMIDs
	has_agent			

Disease [more] [less] *Link to MIM*

Relation	MIM ID	MIM name	PMIDs
agent_in			

Modifiers used in the annotation

Modifier	**Definition**
NOT	*negation of the relation indicated.* Can be used with: participates_in (a biological process) part_of (a complex) located_in (a cellular component) has_function (a molecular function)

	has_part (a domain) Comment: This is used when a PRO term is known to not have the quality indicated by the relation. Example: PR:000000652
contributes_to	*enables in some (possibly unknown) way*. Can be used with: has_function Comment: Only applies to a protein when describing the function of a complex in which it is found. See http://www.geneontology.org/GO.annotation.conventions.shtml#contributes_to. Example: PR:000000682
decreased	*less able relative to normal*. Can be used with: has_function Comment 1: Indicates that the protein performs the function less efficiently that the form indicated in the Relative_to column. Comment 2: It is mandatory to fill the column Relative_to with the PRO ID corresponding to the protein of reference. Example: PR:000002609 relative_to PR:000002605
increased	*more able relative to normal*. Can be used with: has_function Comment1: Indicates that the protein performs the function more efficiently that the form indicated in the Relative_to column. Comment 2: It is mandatory to fill the column Relative_to with the PRO ID corresponding to the protein of reference. Example: PR:000000563 relative_to PR:000002529
altered	*different from the indicated entity, but not in a more-or-less-able way*. Can be used with: has_function part_of Comment: Indicates that the indicated quality differs in some way from the form indicated in the Relative_to column, but “some way” does not include ability. When used with part_of, it indicates that the association of the protein with the complex is unusual in some way (or the complex itself is unusual). Like contributes_to (see below), “altered part_of” is a place holder combination that will disappear once annotations are made directly to the complex (the unusual complex will become an entity directly).

	Example: PR:000000760 relative_to PR:000002604

Relations:

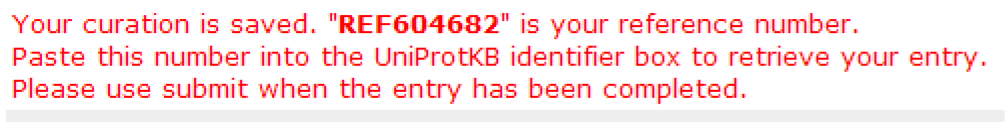
Relation	Ontology	Definition
part_of	GO Component complexes	http://www.obofoundry.org/ro/#OBO_REL:part_of
located_in	GO component subcellular location	http://www.obofoundry.org/ro/#OBO_REL:located_in
has_part	Domain (interpro/pfam)	Inverse of part_of
has_agent	SO mutation causing….	http://www.obofoundry.org/ro/#OBO_REL:has_agent
has_function	GO molecular function	RO_proposed (alt_id OBO_REL:0000031)
participates_in	GO biological process	Inverse relation of has_participant http://www.obofoundry.org/ro/#OBO_REL:has_participant
agent_in	Disease (MIM)	Inverse of has_agent

5. Comment

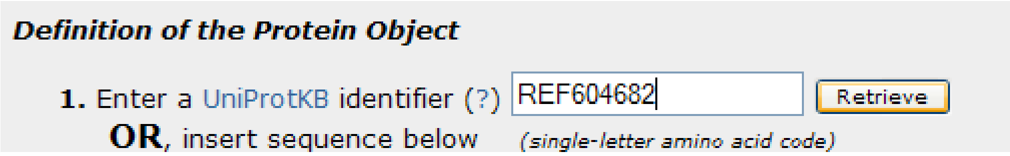
Whenever you need to refer to a PRO node please add the information in this section. For example, if you want to add a new modified form for smad2 isoform 1 (PR:000000468), then it is good that you add in the comment section: "Child of PR:000000468.", or if you are actually adding annotation to an existing PRO term, let's suppose that you want to add annotation to smad2 isoform 1 where the paper describes some attribute of this isoform from mouse. Then after retrieving the corresponding sequence and adding the annotation, please add in the comment: "PRO|PR:000000468;" Any other new comment should go in a separate line.

6.Saving / submitting the annotation:

Save option is to give you the possibility to save your data in case you have not finished and need to return to the annotation later. When you save you are given a REF number and then you can insert this number in the UniProtKB identifier box to retrieve your entry.



Your curation is saved. "**REF604682**" is your reference number.
Paste this number into the UniProtKB identifier box to retrieve your entry.
Please use submit when the entry has been completed.



Definition of the Protein Object

1. Enter a UniProtKB identifier (?) REF604682 Retrieve

OR, insert sequence below *(single-letter amino acid code)*

Submit is used when you are done with the entry. You will still have the same reference number. Please keep for tracking purposes.

What happens next?

Receive an email with the ref number in the subject when your entry is under reviewed

A PRO curator will be assigned to review your entry and create the corresponding PRO node.

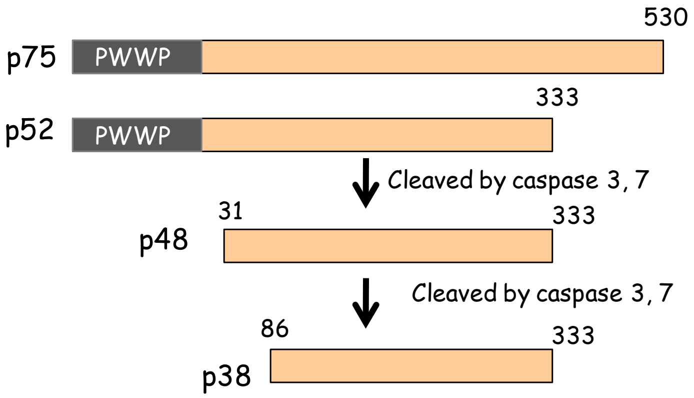
You will receive an email with the PRO ID, and the terms for your final check

Now a real example

Alternative splicing and caspase-mediated cleavage generate antagonistic variants of the stress oncoprotein LEDGF/p75. (PMID: 18708362)

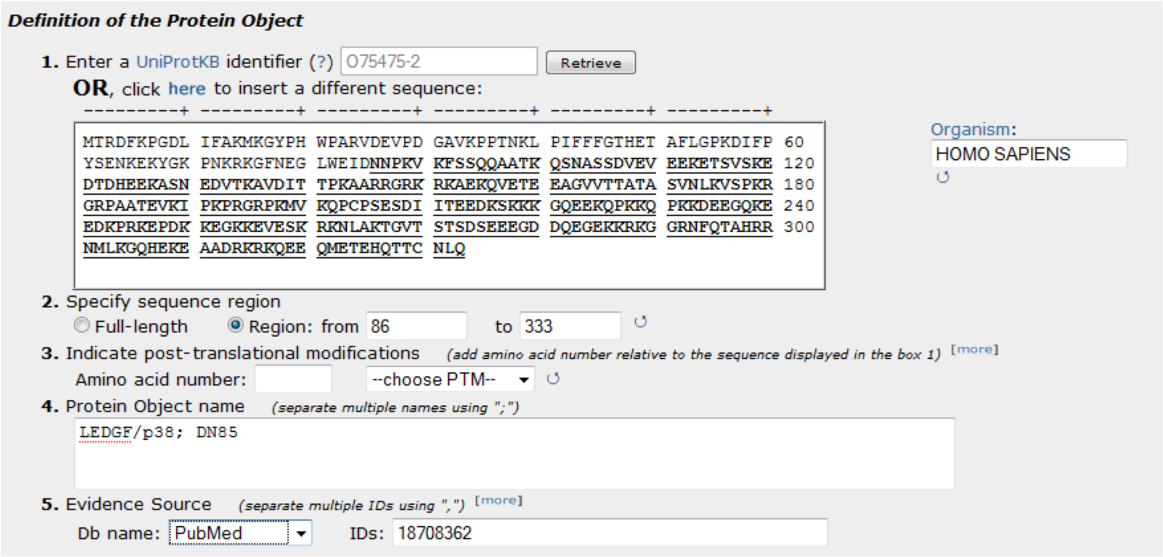
The paper shows that there are two isoforms of the human LEDGF protein, one is p75 the other p52, and describes the proteolytic processing of p52. In the paper there is experimental information for the p75, unprocessed p52, and the two cleaved fragments (p48, and p38).

Therefore in RACE-PRO we should create four entries one for each of these forms. See figure below:



PMID: 18708362

As an example the annotation pertinent for p38 is shown below.



Definition of the Protein Object

1. Enter a UniProtKB identifier (?) O75475-2 Retrieve

OR, click here to insert a different sequence:

```
---------+ ---------+ ---------+ ---------+ ---------+ ---------+
MTRDFKPGDL IFAKMKGYPH WPARVDEVPD GAVKPPTNKL PIFFFGTHET AFLGPKDIFP 60
YSENKEKYGK PNKRKGFNEG LWEIDNNPKV KFSSQQAATK QSNASSDVEV EEKETSVSKE 120
DTDHEEKASN EDVTKAVDIT TPKAARRGRK RKAEKQVETE EAGVVTTATA SVNLKVSPKR 180
GRPAATEVKI PKPRGRPKMV KQPCPSESDI ITEEDKSKKK GQEEKQPKKQ PKKDEEGQKE 240
EDKPRKEPDK KEGKKEVESK RKNLAKTGVT STSDSEEEGD DQEGEKKRKG GRNFQTAHRR 300
NMLKGQHEKE AADRKRKQEE QMETEHQTTC NLQ
```

Organism: HOMO SAPIENS

2. Specify sequence region

○ Full-length ● Region: from 86 to 333

3. Indicate post-translational modifications *(add amino acid number relative to the sequence displayed in the box 1)* [more]

Amino acid number: --choose PTM--

4. Protein Object name *(separate multiple names using ";")*

LEDGF/p38; DN85

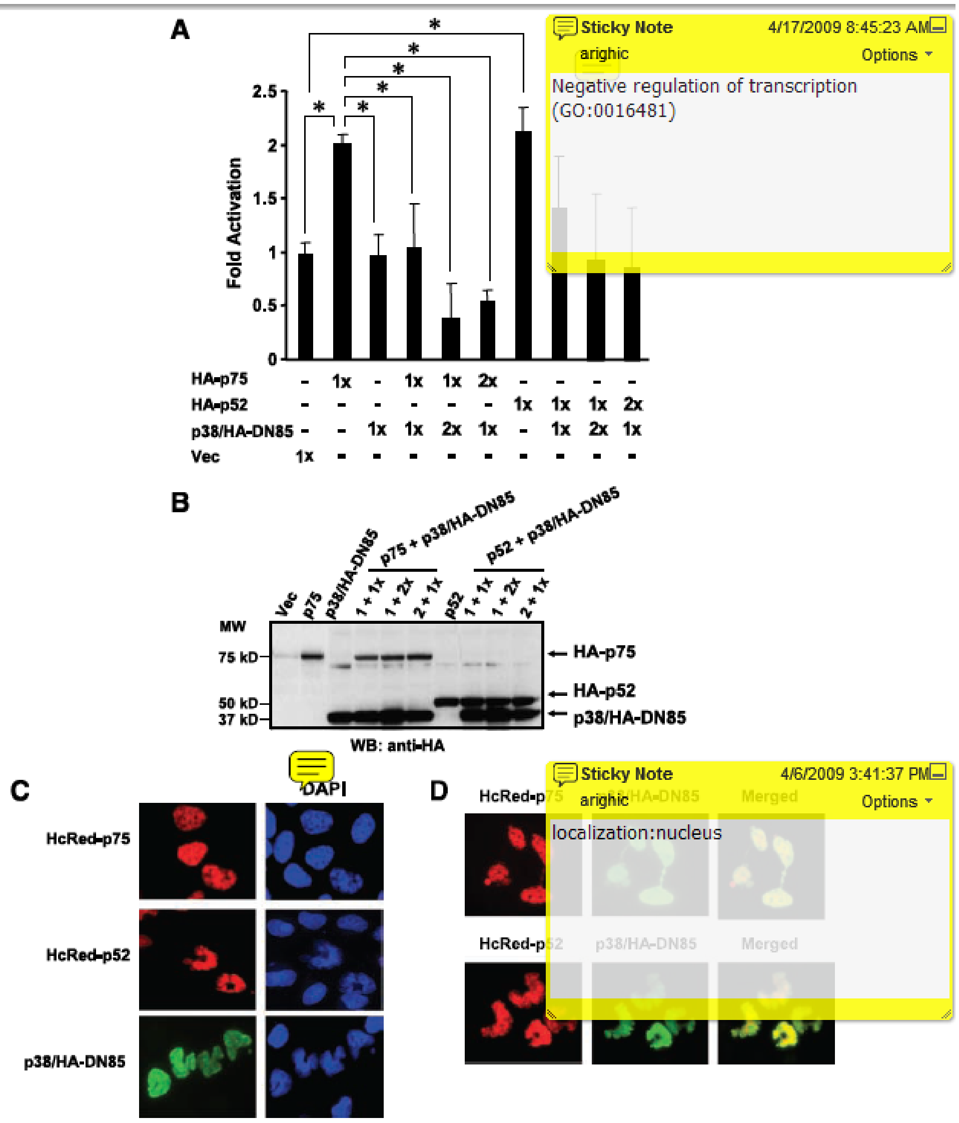
5. Evidence Source *(separate multiple IDs using ",")* [more]

Db name: PubMed IDs: 18708362

Let’s find experimental information for LEDGF/p38.

In the paper you can find the sentence where
“During apoptosis, caspase-3 cleaved p52 to generate a p38 fragment that lacked the NH(2)-terminal PWWP domain and failed to transactivate the Hsp27 promoter in reporter assays.”

So the p38 product lacks the PWWP domain (as determined by sequencing). By doing a search in PFam (http://pfam.sanger.ac.uk/search), PWWP domain corresponds to PF00855 and the name is the same PWWP domain. This negative annotation can be added in the Domain section of the RACE-PRO annotation. See below.



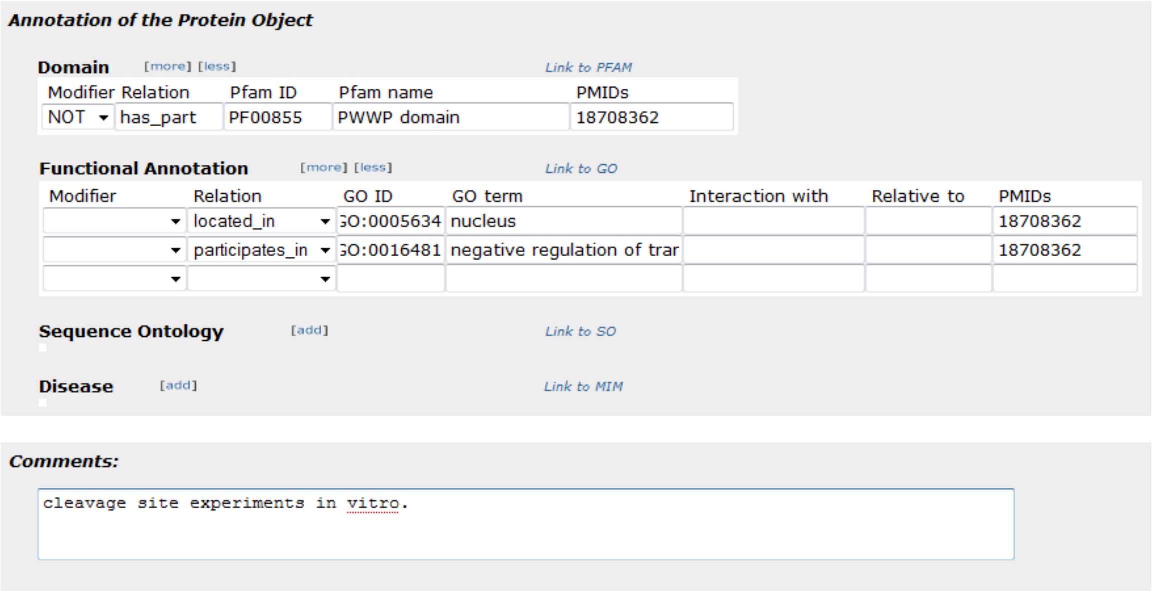
Sticky Note 4/17/2009 8:45:23 AM
arighic Options
Negative regulation of transcription (GO:0016481)

Sticky Note 4/6/2009 3:41:37 PM
arighic Options
localization:nucleus

FIGURE 5. The p38 cleavage fragment interferes with the transactivation potential of LEDGF/p75. **A.** Either 1 μg (1×) or 2 μg (2×) of pCruzHA-LEDGF/p75 or pCruzHA-p52 DNA were cotransfected with 0.2 μg (1×) or 0.4 μg (2×) of pCruzHA-DN85 or empty vector (*Vec*), together with the pGL3-*Hsp27pr*-Luc reporter plasmid. HCT116 cells were lysed after 48 h and assayed for fold induction of luciferase activity. Columns, mean of three independent experiments done in quadruplicate; bars, SD. *, $P < 0.05$, one-way ANOVA with Bonferroni's multiple comparison test (GraphPad Prism). **B.** Corresponding immunoblot showing recombinant protein expression, detected with anti-HA antibody, 48 h after transfection in HCT116 cells. **C.** U2OS cells were transiently transfected with pHcRed-LEDGF/p75, pHcRed-p52, or pCruzHA-p38/DN85 and then grown in coverslips. After 48 h, recombinant protein expression in transfected cells was visualized by fluorescence microscopy. After fixation and permeabilization, HcRed-LEDGF/p75 and HcRed-p52 were visualized directly, whereas HA-p38/DN85 was detected with primary rabbit anti-HA antibody with secondary Alexa 488–labeled goat anti-rabbit antibody. Nuclei were counterstained with 4′,6-diamidino-2-phenylindole (*DAPI*). **D.** Cells were cotransfected with pCruzHA-p38/DN85 (0.5 μg) and pHcRed-LEDGF/p75 (1 μg) or pHcRed-p52 (1 μg), stained with anti-HA antibodies, and visualized by fluorescence microscopy.

The paper shows that the p38 cleaved fragment interferes with the transactivation potential of the full length protein. This could be translated into negative regulation of transcription. Pannel C shows the subcellular localization of the fragment which is nuclear. Then we can search for both GO terms in GO (www.geneontology.org).

Now let’s add all this information in BLOCK 2, the annotation section



Annotation of the Protein Object

Domain [more] [less] *Link to PFAM*

Modifier	Relation	Pfam ID	Pfam name	PMIDs
NOT	has_part	PF00855	PWWP domain	18708362

Functional Annotation [more] [less] *Link to GO*

Modifier	Relation	GO ID	GO term	Interaction with	Relative to	PMIDs
	located_in	GO:0005634	nucleus			18708362
	participates_in	GO:0016481	negative regulation of trar			18708362

Sequence Ontology [add] *Link to SO*

Disease [add] *Link to MIM*

Comments:

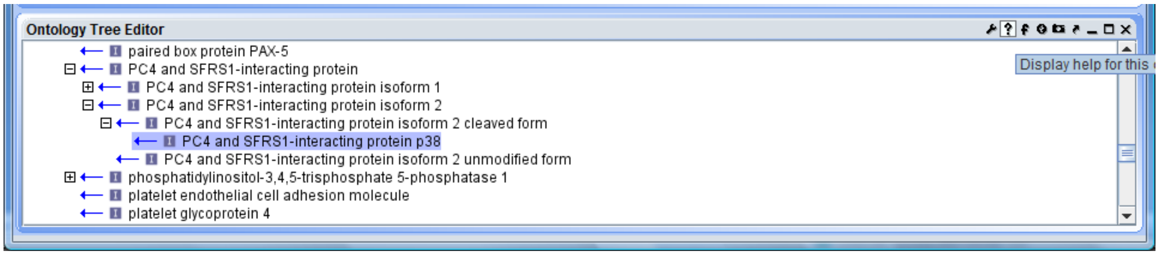
cleavage site experiments in vitro.

After reviewing a PRO stanza is created or updated along with all the related PRO terms.

Stanza created

[Term]
id: PR:000003424
name: PC4 and SFRS1-interacting protein p38
def: "A PC4 and SFRS1-interacting protein isoform 2 cleaved form that is the C-terminal proteolytic product generated by caspase 7 and caspase 3 during apoptosis. Example: UniProtKB:O75475-2, 86-333." [PMID:18708362, YOUR INITIALS]
comment: Category=modification.
synonym: "DN85" EXACT []
synonym: "LEDGF/p38" EXACT []
synonym: "PC4 and SFRS1-interacting protein isoform 2 cleaved 1 " RELATED []
is_a: PR:000025410 ! PC4 and SFRS1-interacting protein proteolytic cleavage product
relationship: derives_from PR:000003422 ! PC4 and SFRS1-interacting protein isoform 2

Your term in the PRO hierarchy



- paired box protein PAX-5
- PC4 and SFRS1-interacting protein
  - PC4 and SFRS1-interacting protein isoform 1
  - PC4 and SFRS1-interacting protein isoform 2
    - PC4 and SFRS1-interacting protein isoform 2 cleaved form
      - PC4 and SFRS1-interacting protein p38
    - PC4 and SFRS1-interacting protein isoform 2 unmodified form
- phosphatidylinositol-3,4,5-trisphosphate 5-phosphatase 1
- platelet endothelial cell adhesion molecule
- platelet glycoprotein 4

Exercise

Now try creating a RACE-PRO entry for one of the forms of the human serase-1b, a splice variant of the TMPRSS9 gene based on the paper corresponding to PMID:16872279. Link to the paper http://www.biochemj.org/bj/400/0551/bj4000551.htm

Some useful text mining tools

1-RLIMS-P (http://pir.georgetown.edu/pirwww/iprolink/rlimsp.shtml), a rule-based text-mining program specifically designed to extract protein phosphorylation information on protein kinase, substrate and phosphorylation sites from the abstracts.

2-IHOP (http://www.ihop-net.org/UniPub/iHOP/), a gene network for navigating the literature

3-GOPubmed (http://www.gopubmed.org/)