

Protein-Centric Connection of Biomedical Knowledge: Protein Ontology (PRO) Research and Annotation Tools

Cecilia N. Arighi¹, Darren A. Natale², Judith A. Blake³, Carol J. Bult³, Michael Caudy⁴,
Alexander D. Diehl⁵, Harold J. Drabkin³, Peter D'Eustachio⁶, Alexei Evsikov³,
Hongzhan Huang¹, Natalia V. Roberts¹, Alan Ruttenberg⁷, Barry Smith⁸, Jian Zhang², Cathy H. Wu^{1, 2}

¹Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA

²Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA

³The Jackson Laboratory, Bar Harbor, ME, USA

⁴Ontario Institute for Cancer Research, Toronto, ON, Canada

⁵Department of Neurology, University at Buffalo, NY, USA

⁶New York University School of Medicine, New York, NY, USA

⁷School of Dental Medicine, University at Buffalo, NY, USA

⁸Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, NY, USA

Abstract. The Protein Ontology (PRO) web resource provides an integrative framework for protein-centric exploration and enables specific and precise annotation of proteins and protein complexes based on PRO. Functionalities include: browsing, searching and retrieving, terms, displaying selected terms in OBO or OWL format, and supporting URIs. In addition, the PRO website offers multiple ways for the user to request, submit, or modify terms and/or annotation. We will demonstrate the use of these tools for protein research and annotation.

1 The Protein Ontology Resources

The Protein Ontology (PRO) is a formal and well-principled Open Biomedical Ontologies (OBO) Foundry ontology for proteins and protein complexes [1]. It is one of the first six ontologies recommended by the OBO Foundry as preferred targets for community convergence, alongside the Gene Ontology (GO). The PRO website (<http://pir.georgetown.edu/pro/pro.shtml>) provides an integrative framework for protein-centric exploration and enables specific and precise annotation of proteins and protein complexes based on PRO. The website functionalities include: i) browsing the ontology while displaying selected data, ii) retrieving a specific branch of the ontology, iii) searching the ontology, mappings and annotations, iv) displaying OBO stanzas for selected terms which can be used into visualization tools such as Cytoscape for an integrated view, and v) downloading selected terms in OWL format for import into an ontology or OWL-aware environment. In addition, each term has a corresponding PRO entry report that links the ontology information, the annotations and the mapping to external resources, therefore displaying all

the information available for that term. For example, a term for a given complex will contain relationships and links to all the individual protein components plus annotation that applies to this complex (**Fig. 1**). PRO identifiers are URIs following the OBO Foundry ID Policy (<http://obofoundry.org/id-policy.shtml>). An example is:

http://purl.obolibrary.org/obo/PR_000000000.

URLs are resolvable, providing information in the web browser and linked data access [2] using Ontobee (<http://ontobee.org>).

PRO allows researchers to explore functional and evolutionary relationships of proteins and protein complexes as well as their higher level organization in pathways and protein networks (**Figs. 1 and 2**). For example, **Fig. 2** shows in a single Cytoscape view that glutaminase 1 has a paralog glutaminase 2 (both share the glutaminase domain as shown in annotation of the parent term), that both are found *E.coli* and *B. subtilis*. It also shows the acetylation of glutaminase 1 and that the active glutaminase 1 is a complex (see corresponding annotation) and it is also observed in both species. A controlled vocabulary is used for annotation and PRO interoperates with GO for PRO complexes.

PRO **Protein Ontology report for entry - PR:000025873** [view OBO stanza](#) [related PRO nodes \(Parent/Child\)](#)

Ontology Information										
PRO ID	PR:000025873									
PRO name	glutaminase 1 tetramer (E. coli K12)									
Synonyms										
Definition	A glutaminase 1 tetramer in Escherichia coli (strain K12). [PRO:CNA, PMID:18459799]									
Comment	Category=organism-complex.									
Hierarchical relationship	Parent: PR:000025872 glutaminase 1 tetramer Children: none has_part PR:000025871 (cardinality="4") glutaminase 1 isoform 1 unmodified form (E. coli K12) (click to see DAG view.)									
Annotation										
	Modifier	Relation	Ontology ID	Ontology Term	Relative_to	Interaction With	Evidence Source	Evidence Code	Taxon ID	Inferred From
Functional Annotation		has_function	GO:0004359	glutaminase activity			PMID:18459799	EXP	83333	

PRO **PRO Hierarchy** (Note that the implicit relationship is *is_a*, whereas ^d indicates *derives_from* relationship.)

14 shown of 26947 records

FANTHER Definition

Synonym Gene PGL PGLC Pfam Isozyme UniProtKB Taxon

expand	sort(m)	sort(m)	glutaminase 1	find	Enter PRO	Category	PIRSF	PMID	EcoCyc
			PR:000018263			amino acid chain			
			PR:000000001			protein			
			PR:000025865			bacterial glutaminase	family	PIRSF005073	18459799
			PR:000022801			glutaminase 1	gene		G6261-MONOMER
			PR:000025882			glutaminase 1 isoform 1	sequence		
			PR:000025084			glutaminase 1 isoform 1 acetylated 1	modification		
			PR:000025083			glutaminase 1 isoform 1 unmodified form	modification		
			PR:000025870			glutaminase 1 isoform 1 unmodified form (Bacillus subtilis)	organism-modification	18459799	
			PR:000025871			glutaminase 1 isoform 1 unmodified form (E. coli K12)	organism-modification	18459799	
			GO:0032991			macromolecular complex			
			GO:0043234			protein complex			
			PR:000025872			glutaminase 1 tetramer	complex		
			PR:000025874			glutaminase 1 tetramer (Bacillus subtilis)	organism-complex	18459799	
			PR:000025873			glutaminase 1 tetramer (E. coli K12)	organism-complex	18459799	

Figure 1. Sample entry report for a glutaminase 1 complex (*upper panel*), the hierarchy of terms related to glutaminase 1 along with selected data (*bottom panel*).

To respond to community needs, the PRO website offers different ways for the user to request, submit or modify terms. A SourceForge tracker can be used to request new PRO terms or modifications of existing ones. Users can submit a request of a few terms or submit a file using a standardized format that can be input into a semi-automatic pipeline for generation of the PRO terms. In addition, domain experts can be actively engaged in the ontology and annotation by submitting to RACE-PRO. This tool allows non-ontologists to author terms and/or annotations. RACE-PRO provides a simple mechanism where the user typically retrieves a protein sequence for the protein form to be described, specifies a sequence region and/or post-translational modification(s) that occurs in the protein form, includes the data source of information (such as PubMed ID), and if need be, adds annotation using controlled

vocabulary (Gene Ontology, MIM, Pfam, Sequence Ontology) (**Fig. 3**). The user is given a reference number to track the annotation and the information is saved internally as a tab delimited file in a format similar to the PRO annotation file (PAF), and checked by a PRO editor. Since most PRO term names and definitions follow a standardized format, a script converts the information therein into PRO terms, by checking for existing terms, and adding parent terms as needed. Once a PRO ID is generated, it is sent to the user along with the PRO term and annotation for a final check and then it is integrated in the PRO release (based on the example in **Fig. 3** two terms were created PR:000026785 and the parent term PR:00002439). We will demonstrate use of these tools to assist protein research and PRO curation.

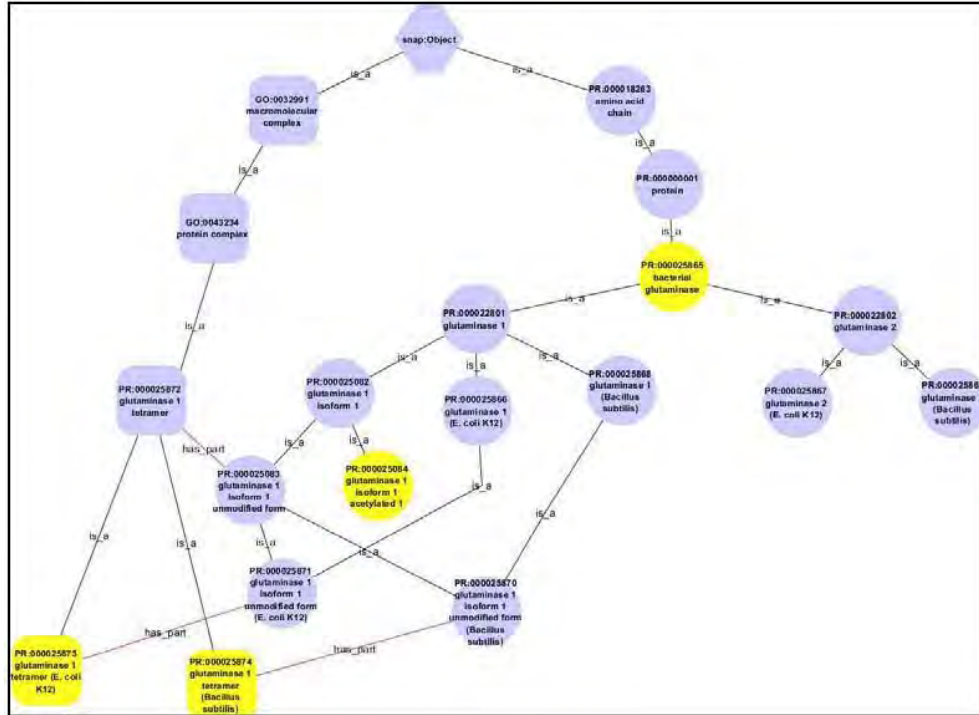


Figure 2. Glutaminase 1-related terms in Cytoscape view.
Reference: circles represent proteins and rounded squares represent protein complexes.

RACE-PRO Rapid Annotation Interface for Protein Ontology (?)

DAF Upload PAF Delete
Save Submit Reset
Tue Jun 7 09:37:25 2011

* Annotator name: _____ E-mail: _____ * Institution: _____
Note: Your e-mail address is for internal use only and will not be shared with third parties.

Definition of the Protein Object

1. * Enter a UniProtKB identifier (?) [P70644-1] [Retrieve] (Related PRO nodes) (REF40201)
OR, click here to insert a different sequence:
[Example: Q21796, Q21796-2, VAR_022776]

RESEVINGSG LGAEHTDILL VFQFLQSSGC TRQEEVLQGR ELFPQAYVEA DLEDELQDQ 60
SQASRSFNG RIEPDSSEGE EIEHNIARHL AQLGDERDHH IQPTLVKRLA ADFPNDLSL 120
EEDSRCLAKA LDEYKTAFAE DQCHDKML I NTHLAKSVA SNAKSLRQV PHTVYFVING 180
HLFSYVRLV RHEED

2. Specify sequence region
Full-length Region: from [1] to [195]

3. Indicate post-translational modifications (add amino acid number relative to the sequence displayed in the box.) [more]
Amino acid number: _____ -choose PTM- [v]

4. * Protein Object name (separate multiple names using ",")
BH3-interacting domain death agonist p15 cleaved form 1; p15

5. * Evidence Source (separate multiple IDs using ",") [more]
Db name: PMID IDs: 9727492

Annotation of the Protein Object

Domain [more] Link to PFM

Functional Annotation [more] [more] Link to GO

Modifier	Relation	GO ID	GO term	Interaction with	Relative to	PMIDs
located_in		GO:000573	mitochondrion			9727492
participates_in		GO:000863	apoptotic mitochondrial c			9727492
located_in		GO:003255	integral to mitochondrial m			9873064
participates_in		GO:000183	release of cytochrome c			9873064

Sequence Ontology [more] Link to SO

Disease [more] Link to MIM

Comments:
active form; this form is the result of Caspase 9 cleavage. The p13 form is the result of Caspase 8 cleavage at position 76.

[Term] PARENT TERM CREATED BASE
id: PR:00002439
name: BH3-interacting domain death agonist isoform
def: "A BH3-interacting domain death agonist protein derived from BH3-interacting domain death agonist by proteolytic cleavage mediated by a caspase to BH3-interacting domain death agonist p15." [PMID comment: Category:modification.
synonym: "BH3-interacting domain death agonist p15" EXACT {}
synonym: "p15 BID" EXACT {}
is_a: PR:000018749 ! BH3-interacting domain death agonist
relationship: derives_from PR:00002258 ! BH3-interacting domain death agonist p15

[Term] TERM CREATED BASED ON PR
id: PR:000026785
name: BH3-interacting domain death agonist isoform
def: "A BH3-interacting domain death agonist isoform derived from BH3-interacting domain death agonist p15 by proteolytic cleavage mediated by a caspase to BH3-interacting domain death agonist p15." [PMID comment: Category:protein sequence.
synonym: "p15" EXACT {}
intersection_of: PR:00002439 ! BH3-interacting domain death agonist p15
relationship: intersection_of: only_in_taxon NCBITaxon:10090 !

Figure 3. RACE-PRO interface. Example of annotation of the BH3-interacting domain death agonist p15 cleaved form.

References

- Natale DA, Arighi CN, Barker WC, Blake JA, Bult CJ, Caudy M, Drabkin HJ, D'Eustachio P, Evsikov AV, Huang H, Nchoutmboube J, Roberts NV, Smith B, Zhang J, Wu CH.: The Protein Ontology: a structured representation

of protein forms and complexes. Nucleic Acids Research 39, D539-545 (2011)

- Ruttenberg A, Rees JA, Samwald M, Marshall MS.: Life sciences on the Semantic Web: the Neurocommons and beyond. Briefings in Bioinformatics 10, 193-204 (2009).