ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

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

Cecilia N. Arighi[1], Darren A. Natale[2], Judith A. Blake[3], Carol J. Bult[3], Michael Caudy[4], Alexander D. Diehl[5], Harold J. Drabkin[3], Peter D'Eustachio[6], Alexei Evsikov[3], Hongzhan Huang[1], Natalia V. Roberts[1], Alan Ruttenberg[7], Barry Smith[8], Jian Zhang[2], Cathy H. Wu[1, 2]

[1]Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA
[2]Protein Information Resource, Georgetown University Medical Center, Washington, DC, USA
[3]The Jackson Laboratory, Bar Harbor, ME, USA
[4]Ontario Institute for Cancer Research, Toronto, ON, Canada
[5]Department of Neurology, University at Buffalo, NY, USA
[6]New York University School of Medicine, New York, NY, USA
[7]School of Dental Medicine, University at Buffalo, NY, USA
[8]Center 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 pf 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.

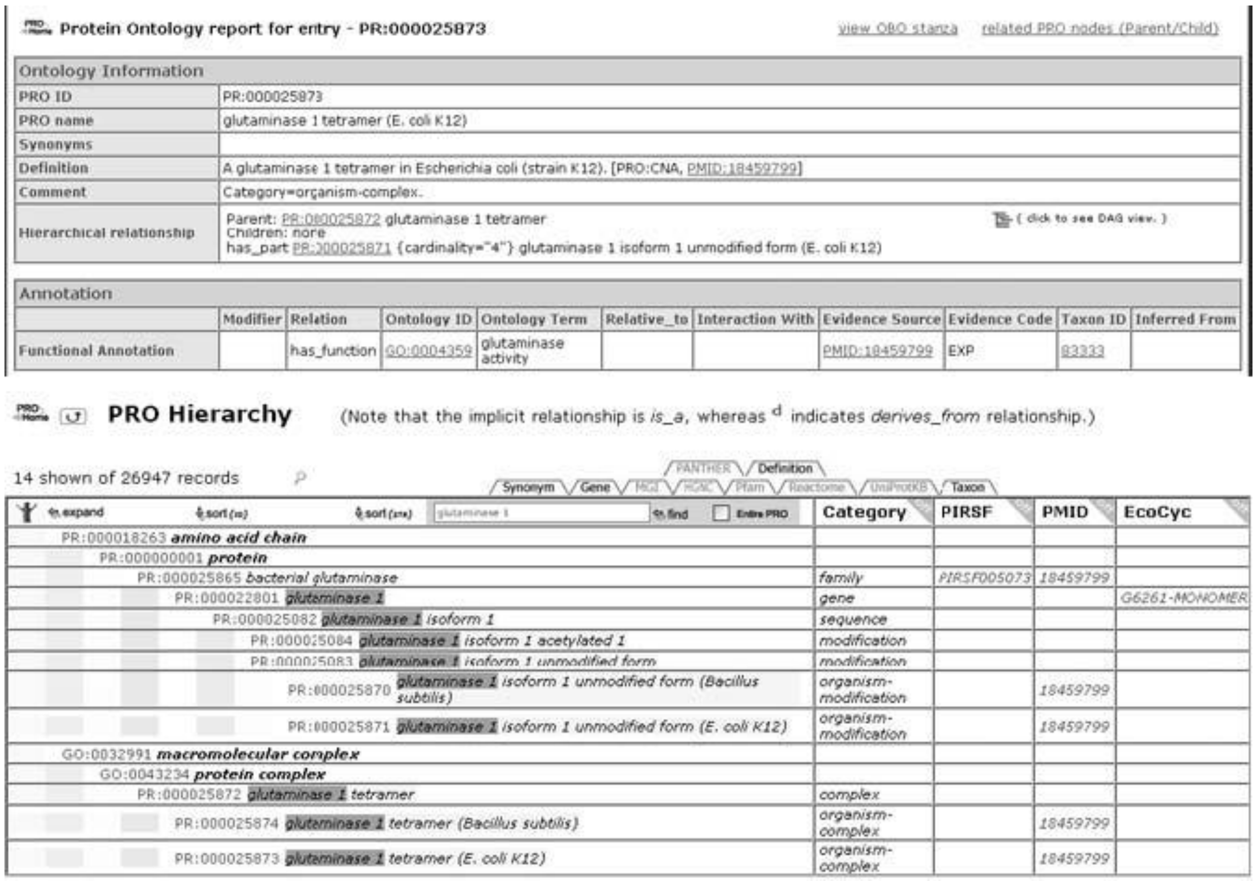

Figure 1. Sample entry report for a glutaminase 1 complex (*upper panel*), the hierarchy of terms related to glutaminase 1 along with selected data (*bottom panel*).

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:000002439). We will demonstrate use of these tools to assist protein research and PRO curation.

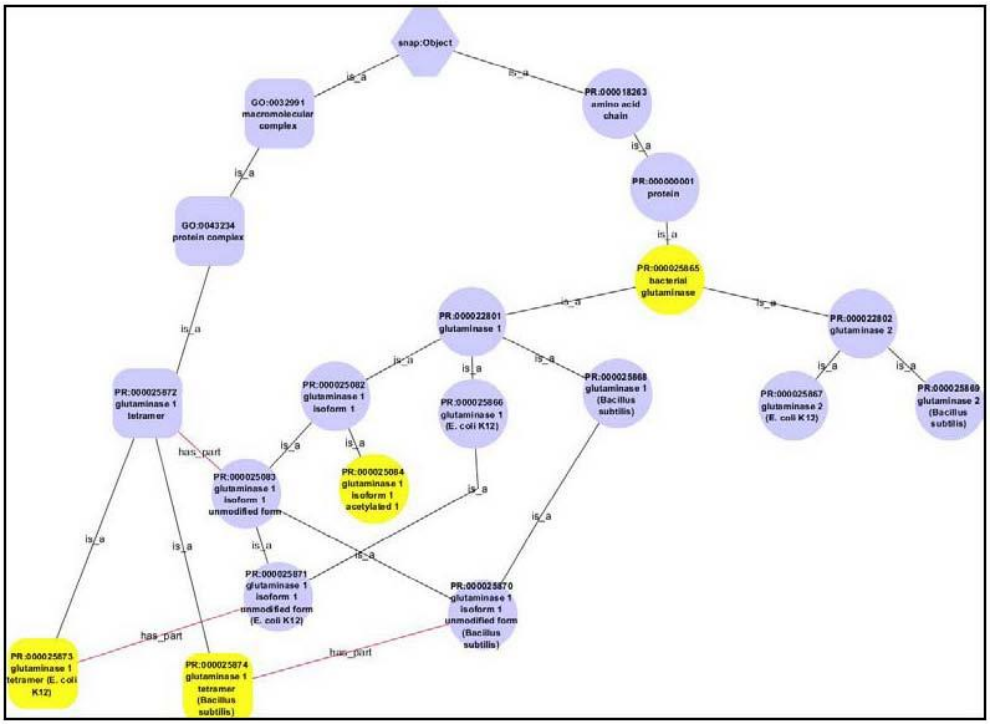

Figure 2. Glutaminase 1-related terms in Cytoscape view.
Reference: circles represent proteins and rounded squares represent protein complexes.

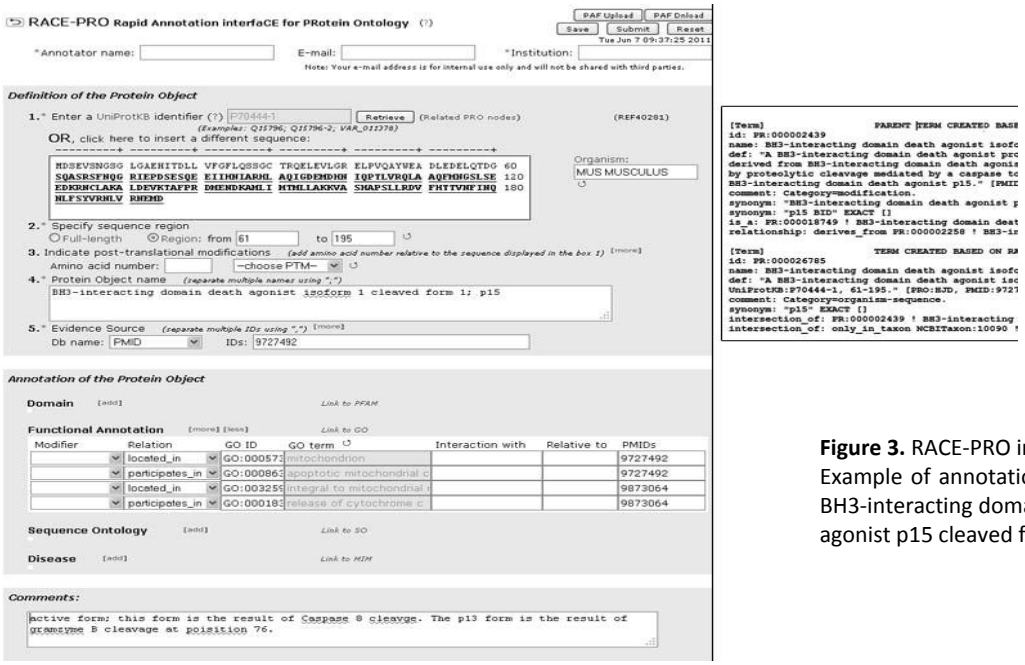

Figure 3. RACE-PRO interface. Example of annotation of the BH3-interacting domain death agonist p15 cleaved form.

References

1. 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)
2. 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).