Framework For A Protein Ontology

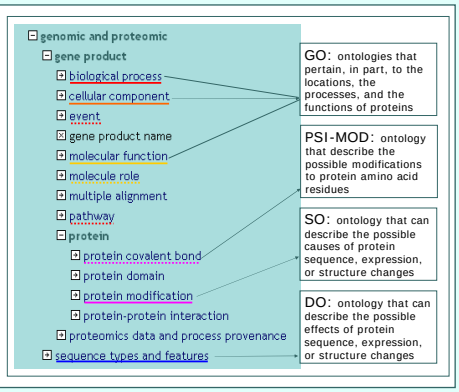
Lai-Su L. Yeh[1], Darren A. Natale[1], Cecilia N. Arighi[1], Winona C. Barker[1], Judith Blake[2], Ti-Cheng Chang1, Zhangzhi Hu[1], Hongfang Liu[3], Barry Smith[4], and Cathy H. Wu[1]

[1]Protein Information Resource, [2]The Jackson Laboratory, [3]Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington, [4]Department of Philosophy, State University of New York at Buffalo

Introduction

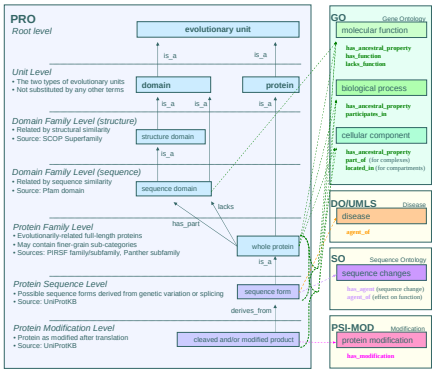
Biomedical ontologies are emerging as critical tools in genomic and proteomic research where complex data in disparate resources need to be integrated. A number of ontologies exist that describe the properties that can be attributed to proteins (See right figure). However, no ontology describes the protein entities themselves and their relationships. For this reason, we have designed a **PR**otein **O**ntology (PRO) that intends to:

- Provide a structure to support formal, computer-based inferences of shared attributes among homologous proteins
- Delineate the multiple protein forms of a gene locus
- Interconnect existing OBO foundry ontologies
- Provide an enabling technology for knowledge discovery



PRO Framework

The components of PRO extend from the classification of proteins on the basis of evolutionary relationships to the representation of the multiple protein forms of a gene (products generated by genetic variation, alternative splicing, proteolytic cleavage, and other post-translational modification). The figure in this panel shows the current working model and a subset of the possible connections to other ontologies.



Blue text boxes: evolutionary component; lavender text boxes: protein form component

PRO Prototype: the transforming growth factor (TGF) beta signaling pathway

PRO connections are first generated by automated extraction of existing annotations from the PIRSF database, for the evolutionary component, and the UniProtKB/Swiss-Prot and MGI entries to get information on the protein forms. Manual curation involves the verification/creation of the protein form nodes and the assignation of the appropriate experimentally verified ontological/controlled vocabulary terms to the sequence forms.

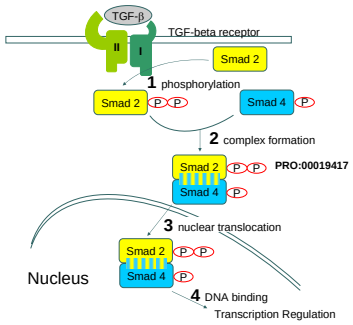
Here we describe the initial development of PRO, illustrated using human and mouse Smad proteins that are essential components of the TGF-beta signaling pathway. These proteins are regulated by phosphorylation and the phosphorylation state dictates cellular location and activity.

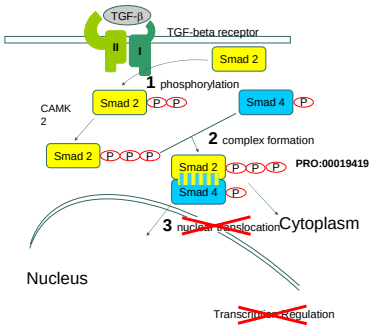
Smad2 in the transforming growth factor beta signaling pathway

The figure below (**1**) depicts the TGF-beta signaling pathway, focusing on the Smad 2 component. The steps shown are preceded by TGF-beta binding to the receptor, and receptor phosphorylation.

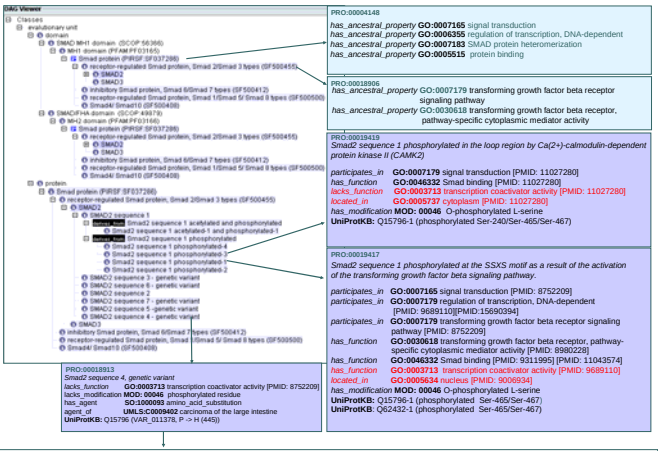
Step 1: Phosphorylation of Smad2 by activated TGF beta receptor I.
Step 2: Complex formation of Smad2 and Smad4.
Step 3: Nuclear import of Smad2:Smad4.
Step 4: Binding of Smad2:Smad4 complex coactivator to responsive element.
The negative regulation by CAMK2 kinase is shown in (**2**)

- (**1**) Activation of Smad2 by TGF-beta receptor
- (**2**) Negative regulation of Smad2 by CAMK2 kinase





PRO can describe the different forms of Smad2 attaching the specific ontology terms. As an example, the two phosphorylated forms depicted in the graph on the left (**PRO:00019417** and **PRO:00019419**) are shown in detail with the differential properties marked in red. In addition, the evolutionary component not only provides the relationship of Smad2 with other Smad protein families (blue boxes), but also may allow to infer properties of homologous proteins in other organisms.



PRO also describes the properties of disease-related genetic variants

Conclusions

- PRO is designed to be a formal, well-principle and extensible OBO Foundry ontology for proteins
- PRO will provide research technology to answer new scientific questions. For example, the transfer of described function/phenotypes of proteins from model organisms to human orthologs may highlight potential candidates to explain human disease
- PRO is expected to create a cycle of improvement for both the ontology and the protein knowledgebases from which the initial information is extracted.

http://pir.georgetown.edu/pro

Contact:
pirmail@georgetown.edu