Protein Ontology Questions and Answers

1) What is the Protein Ontology (PRO)? How does PRO fit into the OBO Foundry set of ontologies?

OBO Foundry ontologies are organized along two dimensions: (1) granularity (from molecule to population); and (2) relation to time (objects, qualities, processes). In terms of this scheme, PRO is a representation of protein objects at the single molecule level of granularity. Specifically, it treats the protein molecules themselves rather than some property of the molecules (such as function, location, molecular weight, etc.). Such properties are instead handled by other ontologies such as GO; PRO provides the objects to which such properties can be attached. PRO encompasses a sub-ontology of proteins based on evolutionary relatedness (ProEvo), and a sub-ontology of the multiple protein forms produced from a given gene locus (ProForm).

2) In what way does the ProEvo part of PRO reflect evolution? There are no terms for ortholog or paralog, nor does the hierarchy seem evolutionarily accurate or complete.

The ProEvo part of PRO is meant to indicate the relatedness of proteins, not their evolution. That is why the relation of a child term to its parent term is *is_a* and not "evolutionarily_derived_from." Nonetheless, the process that resulted in the relatedness (usually reflected in a protein classification) is indeed evolution, and we make every attempt to use evolutionarily-justified classes. Thus, PRO would not group together two proteins that merely share a common function (which might have arose by convergent evolution); instead, PRO would group together two proteins based on evolutionary relatedness, irrespective of function.

3) Why is there no relation between singly-modified forms and multiply-modified forms? Surely the latter once existed as the former.

The important property for each ProForm term is simply that the protein form has been found to exist in nature, not how that form came into existence. Thus, PRO does not take into account the steps leading to a particular form, and there is no hierarchy indicating that, for example, a phosphorylated and ubiquitinated form derived from an original phosphorylated form—the hierarchy is flattened so that each term is a sibling of the other. This is because one cannot assume that the steps leading to a multiply-modified form always occur in the same order.

4) What is meant by the different 'levels of distinction' that are represented in PRO?

The levels of distinction are categories of PRO classes that provide some indication of how PRO is organized. The term "distinction" denotes that the terms at each level can be distinguished from one another as follows:

Family-level distinction: Each PRO term at this level refers to protein products of a distinct gene family arising from a common ancestor. The leaf-most nodes at this level are usually families comprising paralogous sets of gene products (of a single or multiple organisms). For example, smad2 and smad3 both encode proteins that are TGFβ receptor-regulated while smad1, smad5,

and smad9 are all BMP receptor-regulated. Thus, "TGF-beta receptor-regulated smad protein" and "BMP receptor-regulated smad protein" are terms denoting distinct families. Note that this level collectively refers to any such grouping at any level of similarity. For example, the two families indicated above can merge into a "receptor-regulated smad protein" class and further merge (with the protein products of smad4, smad6, and smad7) into the "smad protein" class. No merging occurs beyond what can be done to account for the evolution of the entire, full-length protein. That is, all proteins that can trace back to a common ancestor over the entire length of the protein are part of the same family.

Gene-level distinction: Each PRO term at this level refers to the protein products of a distinct gene. For example, "smad2" and "smad3" are two different genes, even though they are paralogs, and therefore have two different PRO entries at the gene level of distinction. The protein products of all alleles of what is recognized as smad2 in humans and what is recognized as smad2 in mouse thus fall under this single term. Thus, a single term at the gene-level distinction collects the protein products of a subset of orthologs for that gene (the subset that is so closely related that its members are considered the same gene). Gene-level distinction is the leaf-most node of the ProEvo part of PRO.

Sequence-level distinction: Each PRO term at this level refers to the protein products with a distinct sequence upon initial translation. The sequence differences can arise from different alleles of a given gene, from splice variants of a given RNA, or from alternative initiation and ribosomal frameshifting during translation. One can think of this as a mature mRNA-level distinction. For example, smad2 encodes both a long splice form and a short splice form. The protein products from each isoform are separate PRO terms. Sequence-level distinction is the first (parent-most) node of the ProForm part of PRO.

Modification-level distinction: Each PRO term at this level refers to the protein products derived from a single mRNA species that differ because of some change (or lack thereof) that occurs after the initiation of translation (co- and post-translational). This includes sequence differences due to cleavage and chemical changes to one or more amino acid residues. For example, the long isoform of smad2 can either be unmodified or be post-translationally modified to contain phosphorylated residues. Modification-level distinction is the leaf-most node of the ProForm part of PRO.

5) What is the difference between an 'isoform' and a 'sequence variant'?

In the terminology used by PRO, a protein is called 'isoform' when its primary sequence differs from that of another protein encoded by the same gene in a given organism due to alternative splicing or start site selection. A protein that differs due to genetic differences is termed 'sequence variant'.

6) What is the justification for isoform terms in PRO that cross-reference to more than one species? For example, how can an isoform from human be considered the "same" as an isoform from mouse, especially if the sequences differ?

For ease of reference we categorize such isoforms as ‘ortho-isoforms’ (orthologous isoforms). These are isoforms--encoded by orthologous genes--that are believed to have arisen prior to speciation and divergence of the primary sequence. That is, ortho-isoforms were true alternative isoforms (as defined above) in a common ancestor, and quite likely functionally equivalent.

7) Why is each modification of a protein given a different term in PRO?

Proteins in PRO are distinguished on the basis of chemical composition. That is, instances of a protein with a given chemical composition are classed with others with the same composition, and sibling classes will always differ in composition. For example, if a protein obtains a phosphate group (thereby modifying its composition), it is given a different PRO term. Note that instances of modified forms don’t always have the same molecular function as instances of the parent form.

8) What does it mean that PRO is species-neutral?

Species-neutrality indicates that proteins from any species can be included, and that some of the terms apply to proteins in one species, some apply to those in another species, and some apply to several species at once. It does not mean that only proteins found universally to all species are included.

9) Some terms in PRO were obtained via large-scale processing of information from external resources. How can these be identified?

In terms of content and quality, they cannot be distinguished. Large-scale processed information was at all steps manually monitored for quality and accuracy, and the wording of definitions and other standardization was based on what were used for existing manual entries.
Large-scale processed entries contain the phrase "Flag=automatic." On the web site, perform the following search: “Comment” = “Automatic” to get the list.