



The left panel shows the current framework with the different levels of distinctions.

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 PRO framework description

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.

The right panel denotes the different ontologies or databases that are used to either (i) define a protein class, or (ii) annotate a property of an entity of the class. For example, at the modification-level, *smad2* isoform 1 phosphorylated form can be defined as the intersection of PRO *smad2* isoform1 and MOD:00696 phosphorylated residue, using the relation `has_modification`. One such phosphorylated form, *smad2* isoform 1 phosphorylated 1, participates_in signal transduction. Such annotations are provided in the protein ontology association file (PAF) file.

The arrows indicate to what levels these ontologies can be applied