BMC Bioinformatics

Proceedings

Open Access

Framework for a Protein Ontology

Darren A Natale*†1, Cecilia N Arighi†1, Winona C Barker1, Judith Blake2, Ti-Cheng Chang1, Zhangzhi Hu1, Hongfang Liu3, Barry Smith4 and Cathy H Wu1

Address: 1Department of Biochemistry and Molecular & Cellular Biology, Georgetown University Medical Center, 3300 Whitehaven St., NW, Washington, DC, 20007, USA, 2The Jackson Laboratory, 600 Main St., Bar Harbor, ME, 04609, USA, 3Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Room 176, Building D, Washington, DC, 20057, USA and 4Department of Philosophy, State University of New York at Buffalo, Park Hall, Buffalo, NY, 12460, USA

Email: Darren A Natale* - dan5@georgetown.edu; Cecilia N Arighi - cna5@georgetown.edu; Winona C Barker - wb8@georgetown.edu; Judith Blake - jblake@informatics.jax.org; Ti-Cheng Chang - tc245@georgetown.edu; Zhangzhi Hu - zh9@georgetown.edu; Hongfang Liu - hl224@georgetown.edu; Barry Smith - phismith@buffalo.edu; Cathy H Wu - wuc@georgetown.edu

* Corresponding author †Equal contributors

from First International Workshop on Text Mining in Bioinformatics (TMBio) 2006
Arlington, VA, USA. 10 November 2006

Published: 27 November 2007

BMC Bioinformatics 2007, **8**(Suppl 9):S1 doi:10.1186/1471-2105-8-S9-S1

This article is available from: http://www.biomedcentral.com/1471-2105/8/S9/S1

© 2007 Natale et al; licensee BioMed Central Ltd.

Abstract

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 describe properties that can be attributed to proteins. For example, protein functions are described by the Gene Ontology (GO) and human diseases by SNOMED CT or ICD10. There is, however, a gap in the current set of ontologies – one that describes the protein entities themselves and their relationships. We have designed the PRotein Ontology (PRO) to facilitate protein annotation and to guide new experiments. 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 modifications). PRO will allow the specification of relationships between PRO, GO and other ontologies in the OBO Foundry. Here we describe the initial development of PRO, illustrated using human and mouse proteins involved in the transforming growth factor-beta and bone morphogenetic protein signaling pathways.

Background

Ontology-based methodologies for data integration promote precise communication between scientists, enable information retrieval across multiple resources, and extend the power of computational approaches to perform data exploration, inference and mining [1-3]. The Open Biomedical Ontologies (OBO) library [4] is an umbrella for ontologies shared across different biological and medical domains. There is, however, a gap in the current OBO library of ontologies – a protein ontology that defines proteins, protein classes, and their relationships. Filling this gap will support machine reasoning over the shared features of related proteins and will foster the interconnection between ontologies that describe protein attributes. Here we present the framework and initial development of the Protein Ontology (PRO) [5] that describes types of proteins and the relations between them. PRO will concentrate on (a) an ontology of proteins based on evolutionary relatedness (ontology for protein evolution) and (b) an ontology of the multiple protein forms produced from a given gene locus (ontology for protein forms).

Biomedical ontologies

Like standardized measures and rules of syntax, ontologies bring the benefits of synchronization. Thus, for an ontology to be of maximal value, it is crucial to ensure that for each domain of inquiry there is community convergence on a single ontology. The Open Biomedical Ontologies (OBO) provides a resource where biomedical ontologies are made available in a standard format that allows systematic updating and versioning on the basis of community feedback. Currently, there are nearly 60 ontologies distributed through the OBO web site, spanning domains from anatomy (e.g., Mouse adult anatomy) to ethology (Loggerhead nesting), and from gene and gene product features (Sequence Ontology and Gene Ontology) to phenotypic qualities knowledge (Disease Ontology).

The OBO Foundry [6], a consortium formed by a subset of developers of OBO ontologies, has outlined a set of principles specifying best practices in ontology development that are designed to foster interoperability and ensure a gradual improvement of quality and formal rigor. Ontologies in the OBO Foundry are required to be well-documented, to adopt a common formal language, and to be developed in a collaborative manner. The following summarizes several candidate OBO Foundry ontologies that are related to PRO.

Gene Ontology (GO)

The Gene Ontology (GO) is by far the most widely used ontology in any discipline [7]. It aims to formalize the expression of information about biological processes, molecular functions, and cellular components through a controlled vocabulary structured in three mutually independent hierarchies. The GO has been used to annotate the genes of humans and a variety of model organisms, and it has facilitated both in-depth understanding of biology within a single organism and comparison of biological processes across multiple organisms.

Sequence Ontology (SO)

The Sequence Ontology (SO) provides a rich set of terms, relations, and definitions for genome and chromosome annotation [8]. A subset of SO terms addresses the consequences of gene mutation on protein products; for example, whether the mutation decreases or eliminates protein activity.

Disease Ontology (DO)

Still a work in progress, DO [9] is designed to be a controlled medical vocabulary to facilitate the mapping of diseases and associated conditions to medical coding systems such as the ICD10 [10] and SNOMED CT [11], and to other vocabularies within the Unified Medical Language System (UMLS) [12].

Other protein-related ontologies

The PSI-MOD ontology [13] has a comprehensive collection of terms for annotations that describe various types of protein modifications, including cross-links and pre-, co- and post-translational modifications. PSI-MOD is partly constructed using RESID [14] terms, a controlled vocabulary for defining modification features of protein entries in the UniProt Knowledgebase (UniProtKB) [15]. PSI-MI [16] and the INOH Event Ontology (EO) [17] are ontologies that describe some of the types of protein interaction events. Finally, both the INOH Molecule Role Ontology [18] and Reactome [19], contain molecular functional group names, abstract molecule names and concrete molecule names manually collected from literature. The structure of each of the other protein-related resources described above aligns well with PRO.

Two other ontologies have been designed for database integration or annotation. Protein Ontology (PO) [20] includes terms and relationships to describe attributes of individual protein forms (such as physicochemical properties), while the Proteomics Process Ontology (ProPreO) [21] enables a detailed description of proteomics experimental processes and data. Neither ontology includes the protein forms themselves.

Protein Ontology development

The development of PRO will proceed by taking a pragmatically motivated approach to populating classes, relations, and annotations. We start with an initial set of types using existing, complementary, curated protein classification resources. Relations between these types are defined following the methodology of the OBO Relations Ontology [22]. Connections to other ontologies are used to formulate annotations of PRO classes. Finally, the results are subjected to manual validation by experts.

An overview of PRO is provided in Figure 1. For brevity, we refer to the protein evolution component as ProEvo and the protein forms component as ProForm.

Protein Evolution component (ProEvo)

The diverse proteins we find today in living organisms can be grouped into protein families, each member of which

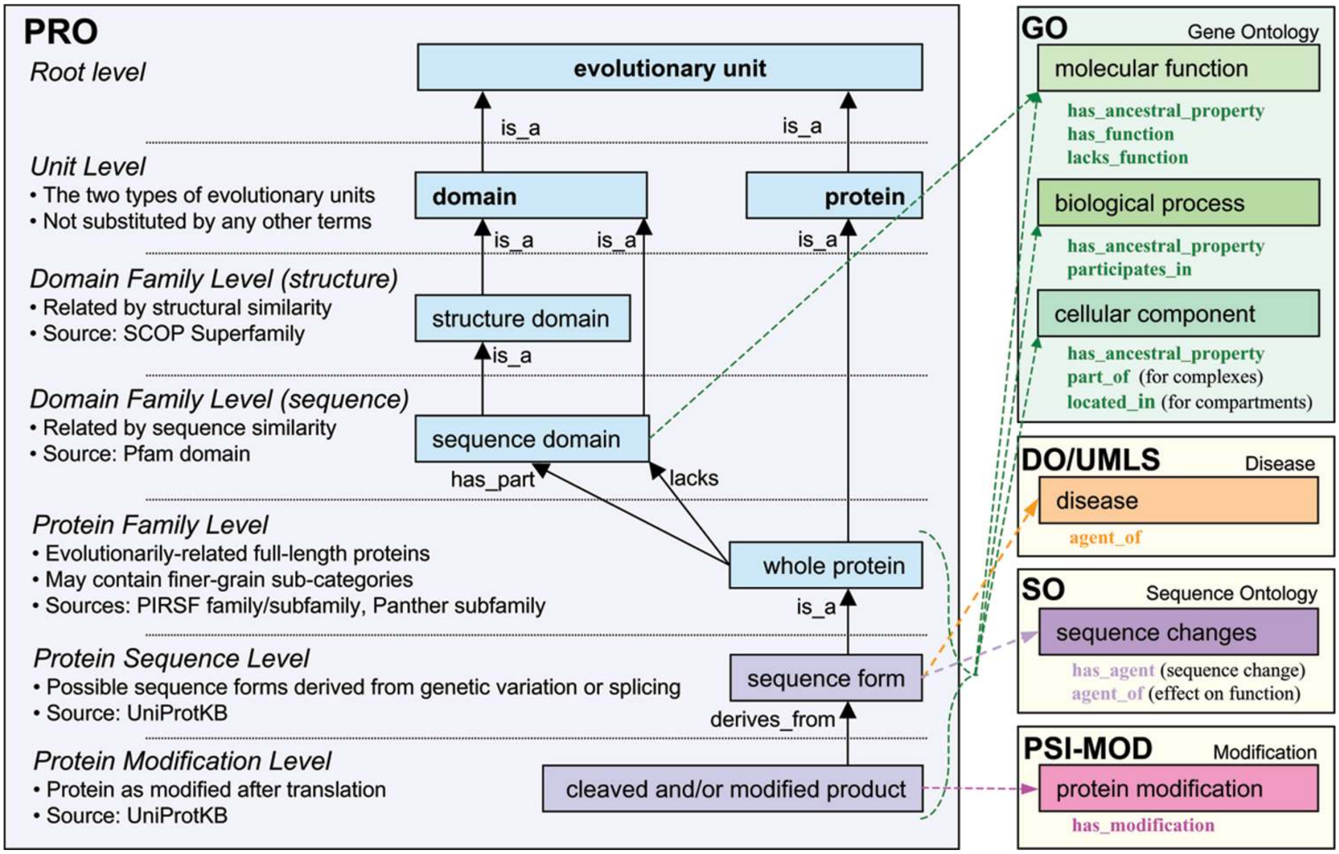

PRO
Root level
evolutionary unit
is_a
is_a
Unit Level
• The two types of evolutionary units
• Not substituted by any other terms
domain
protein
is_a
is_a
is_a
Domain Family Level (structure)
• Related by structural similarity
• Source: SCOP Superfamily
structure domain
is_a
Domain Family Level (sequence)
• Related by sequence similarity
• Source: Pfam domain
sequence domain
has_part
lacks
Protein Family Level
• Evolutionarily-related full-length proteins
• May contain finer-grain sub-categories
• Sources: PIRSF family/subfamily, Panther subfamily
whole protein
is_a
Protein Sequence Level
• Possible sequence forms derived from genetic variation or splicing
• Source: UniProtKB
sequence form
derives_from
Protein Modification Level
• Protein as modified after translation
• Source: UniProtKB
cleaved and/or modified product
GO
Gene Ontology
molecular function
has_ancestral_property
has_function
lacks_function
biological process
has_ancestral_property
participates_in
cellular component
has_ancestral_property
part_of (for complexes)
located_in (for compartments)
DO/UMLS
Disease
disease
agent_of
SO
Sequence Ontology
sequence changes
has_agent (sequence change)
agent_of (effect on function)
PSI-MOD
Modification
protein modification
has_modification

Figure 1
PRO protein ontology overview. The figure shows the current (partial) working model and a subset of the possible connections to other ontologies. Blue text boxes:ProEvo component; lavender text boxes: ProForm component.

derives from a common ancestor. Families have built up over time by copying events (speciation or gene duplication), followed by divergence of the copies from each other. This expansion of a protein family can be represented as a bifurcating tree: each bifurcating node represents the copying of an ancestral sequence. These ancestral sequences are now extinct, but they are inferred from the sequences we observe today. Despite the passage of millions of years of divergence, members of each family still share recognizable similarities. It is therefore often possible to infer certain properties of the ancestral protein, such as function, based on the recognizable similarities of its modern descendants.

During the process of protein evolution, there are portions of proteins – called *domains* – that are usually copied in their entirety, presumably because they represent a minimal functional unit. A *protein* comprises one or more domains, usually with additional sequences connecting and surrounding them. Note that using our definition of domain, some domains have never combined as modules with another domain (at least as observed thus far). Proteins with similar domain architecture (that is, the same combination of domains in the same order) are said to be *homeomorphic*. In the case of single-domain proteins, the evolutionary history is identical to (or is a subtree of) that of the domain itself. However, the evolutionary history of multi-domain proteins is more complex: it can only be represented by a single tree as far back as the earliest ancestor that contained the same architecture. Prior to that, one must look to the histories of the constituent domains.

The relationship between a protein and each of its constituent domains can be modeled using the *has_part* relationship already defined in the OBO Relation Ontology. The relationship is most obvious for multi-domain proteins, but it also holds for single-domain proteins.

One complication is that domains within a multi-domain protein can be lost in one or more lineages (e.g., [23,24]). This means that a *has_part* relationship to this domain that obtains for the parent class will not obtain for the child class. Therefore, we will use a *lacks* relationship type to describe evolutionary loss in the child lineage [25].

The GO molecular function ontology organizes function classes from the general (at the top of the hierarchy) to the specific (at the leaf nodes). In contrast, the hierarchy of ProEvo classes is based on evolutionary relatedness, organized from the distantly-related (at the top of the hierarchy) to the more closely related (at the leaf nodes). In many cases the functional and evolutionary classes will overlap. However, consider the case of erythrocyte membrane protein band 4.2 (EPB4.2). This protein is a major component of the red blood cell membrane skeleton [26] that was co-opted from an ancestral protein-glutamine gamma-glutamyltransferase [27], but subsequently lost the ancestral function [28]. In the GO molecular function ontology, the appropriate association for EPB4.2 is "constituent of cytoskeleton" (GO:0005200). For PRO, its parent is "protein-glutamine gamma-glutamyltransferase." The evolutionary relationship between the human and mouse versions of EPB4.2 and protein-glutamine gamma-glutamyltransferase is represented schematically in Figure 2. The difference in function is not due to gain or loss of specific sections of protein (domains), since all four proteins share end-to-end similarity and common domain architecture. However, two of the residues of the catalytic triad of protein-glutamine gamma-glutamyltransferase [29] are changed in EPB4.2 (data not shown).

Populating ProEvo classes: Resources

Several resources exist that group proteins according to function, sequence or structure-based relatedness. We use four of these resources to guide the initial construction of PRO. Together, these resources represent all of the basic elements of a protein evolutionary ontology outlined above. They provide the set of classes that are most important for one of the primary tasks we wish to accomplish with the evolution component: reliably using experimental data from other organisms to understand human genes. Moreover, each of these four resources has been curated by expert biologists to ensure quality. For clarity, in the description of these resources, we refer to the sets of proteins as "groups" or "families" or "clusters," and the name given to the set as the "class." The section below lists each resource according to the evolutionary relationships for which each approach is most appropriate, from the most distant to the closest.

Structure-based clusters with remote domain homology: SCOP

SCOP (Structural Classification of Proteins) [30] is arranged hierarchically into four levels: *class*, *fold*, *superfamily* and *family*. Homology (common ancestry) can be asserted for proteins in the same family on the basis of sequence data alone and for proteins in the same superfamily on the basis of three-dimensional (3D) structure data. Proteins in different superfamilies in the same fold group or class have similarities in 3D topology but do not necessarily have a common ancestor. Therefore, only the SCOP superfamily and family data are relevant for the purposes of PRO, with the former defining remote homology (shared ancestry that diverged in the distant past) and the latter defining close homology (shared ancestry that diverged in the more recent past).

Sequence-based clusters with close domain homology: Pfam

Pfam domain families [31] are comparable to SCOP families. However, Pfam contains domain definitions even in the absence of structure information; thus, Pfam represents a superset of SCOP families. Accordingly, we will use Pfam domain families in place of SCOP families to represent the "close" level of evolutionary relatedness for domains.

Clusters of homologous proteins: PIRSF

The PIRSF family classification system provides protein classification from superfamily to subfamily levels in a network structure to reflect the evolutionary relationship between sets of whole proteins and between whole proteins and domains [32]. The primary PIRSF classification unit is the *homeomorphic family*, whose members are *homologous* (evolved from a common ancestor) and *homeomorphic* (sharing full-length sequence similarity and a common domain architecture). Basing classification on whole proteins allows annotation of family-specific biological functions, biochemical activities, and sequence features, while an understanding of the domain architecture of a protein provides insight into its general functional and structural properties as well as into complex evolutionary mechanisms.

Functionally-diverged subfamilies: PANTHER

A PANTHER subfamily [33] is defined as a monophyletic group of proteins that have distinct functions as compared to other monophyletic groups in the same protein family. These functional differences can derive from gain and loss of additional domains or from changes in the protein sequence.

Populating ProEvo classes: Mechanism

The initial ProEvo classes will derive from the curated protein clustering resources described above. How one class relates to another consequently resolves to how each cluster relates to another, and the problem condenses to a simple mapping exercise. The relationships between SCOP clusters and Pfam clusters already exist, as do the relationships between Pfam, PIRSF, and PANTHER. To facilitate updates and tracking between these initial resources and ProEvo classes, we will use both PRO accessions and IDs, similar to the system used by UniProt [15].

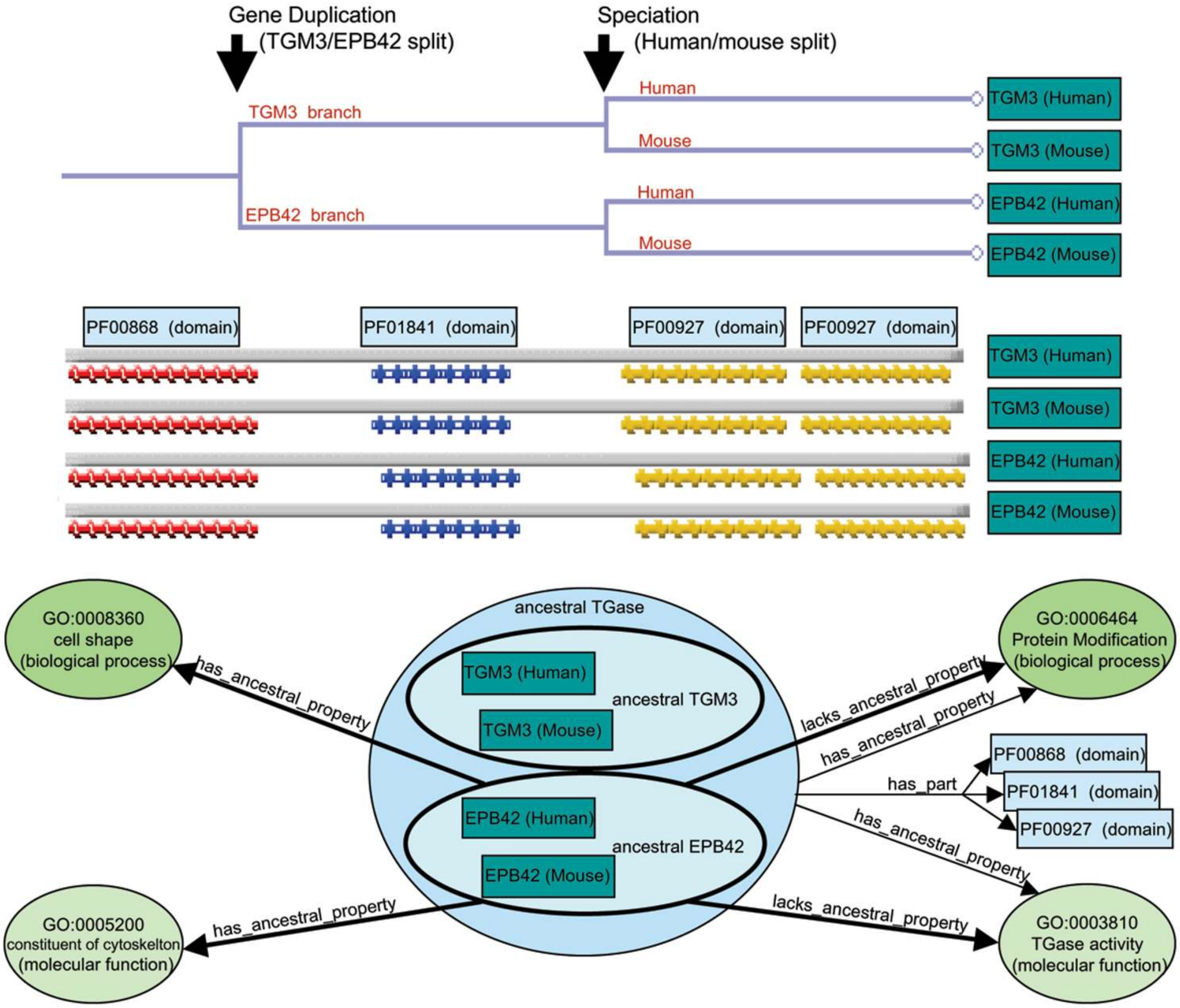


Figure 2

Schematic representation of the evolutionary relationship between human and mouse versions EPB42 and TGM3. Top panel: protein-glutamine gamma-glutamyltransferase (TGM3) and erythrocyte membrane protein band 4.2 (EPB42) are descended from a common ancestor with glutamyltransferase activity. Middle panel: All four proteins share a common domain arrangement. Bottom panel: PRO and GO connections. TGM3 and EBP42 are descended from an ancestral transglutaminase (TGase) represented by the large circle. This protein evolved into the ancestral forms of TGM3 and EPB42, represented by the ovals. All descendants of the ancestral TGase comprise the parts indicated by blue boxes. Except where indicated, all descendants of the ancestral TGase have transglutaminase activity and are involved in protein modification. However, the ancestral EPB42 lacks these attributes, and instead acquired new attributes not shared by the ancestral TGM3.

Thus, whenever possible, each PRO class will have an incremented number as its accession (e.g., PRO:00000001) and a source-database cross-reference (e.g., PRO:PIRSF000001).

Updating ProEvo classes

Once the initial mapping is done – and the classes and relationships are verified – the composition of the underlying clusters and how *they* interrelate will not be of consequence to PRO except as a source of additional nodes. That is, source database changes need not be reflected in the ontology. Consider the example of the hexokinase family of proteins, which includes xylulokinase [34]. Suppose the initial population of PRO classes yields *xylulokinase is_a hexokinase*, and the source database is subsequently modified such that the original xylulokinase family is renamed to ketopentose kinase after adding ribulokinases. The original relation still holds even though the xylulokinase family no longer exists, so the PRO xylulokinase class does not get deleted. Instead, a new level – ketopentose kinase – could be inserted between xylulokinase and hexokinase.

Protein Forms component (ProForm)

A number of different protein forms can be derived from a single gene. Protein databases typically represent only one reference sequence for a gene product, and do not have separate entries for mutations that can give rise to disease, for different forms that arise through variations in splicing, or for post-translational modifications. For example, cleavage of a signal peptide is needed for protein secretion. Also, specific residues can be covalently modified with a variety of chemical moieties. Some proteins engage in cyclic processes that involve, for example, phosphorylation and dephosphorylation. These various modified forms of a given gene product are critical to making precise annotation. For example, many diseases are not caused by the "normal" protein, but by a genetic variant. Also, a protein can activate a process when in its phosphorylated form, but inhibit that same process when not. Such nuances are not possible with the existing ontologies. Therefore, PRO allows for the definition of sequence forms arising from genetic, splice, and translational variation, and from post-translational cleavage and modification.

Relationships between protein forms will be simple and direct and make use of existing relations whenever possible. We will use OBO's Relations Ontology as a source of well-defined relationships, adding further relationships on the basis of need. For example, it is biologically reasonable to say that the product of a post-translational modification is *modified_from* the initial protein. However, using such a relationship adds complexity to the system and hinders the possible interconnections with other ontologies. In fact, *modified_from* is just a more specific way of asserting that "new entity *created_from* old entity" or "new entity *derives_from* old entity." The three relations convey identical ideas, but the latter is already part of the core set of relations [22]. Note, however, that this existing relation does not accurately describe the relationship between two variations of the same gene product, nor does the *is_a* relation. Therefore, we use a new relation *variant_of* for this situation.

Populating ProForm classes: Resources

Both the richness and the usefulness of an ontology stem from the diversity and comprehensiveness of its classes. Accordingly, we intend to capture the diverse forms that a protein can take. UniProtKB/Swiss-Prot contains information on sequence variants due to mutation, alternative splicing, or protein cleavage, and on post-translational modification. These data, found within the controlled vocabulary of FT (feature) lines or free text of CC (comment) lines, have been used to populate the appropriate classes. Other sources of data from which information can be computationally extracted include MGI [35] and iProClass [36].

Populating ProForm classes: Mechanism

We have developed a parser to transform information from the sources indicated above into OBO format nodes and relationships. The parser captures experimentally verified biological entities, ignoring those labeled as "by similarity," "potential," or "probable." There are three kinds of entities considered by the parser: isoforms, variants, and cleavage and modification products. For example, post-translational modification nodes are automatically populated from UniProtKB based on the FT field and from iProClass based on the PIR Feature and Post Translational Modification fields. Automatically-populated ProForm terms are verified by a curator and edited using OBO-EDIT. Additional terms are added as necessary after curator review of the literature.

Connecting to - and between - other ontologies

Several ontologies – notably, GO and DO – are pertinent to protein annotation in that the goal is to annotate gene products. However, the current representation of proteins in databases is not amenable to direct connection of annotation to the appropriate protein forms themselves. The development of PRO provides this necessary intermediary for connections *between* these other ontologies. For example, the logical connection between the process term X and the disease term Y is the protein form Z.

ProEvo connections

Though it is most logical and accurate to connect the attributes available from other ontologies to specific terms in the ProForm component, it is nonetheless useful to make connections to terms of the ProEvo component as well. Doing so provides the ability to reason across species – for example, to apply knowledge obtained from a mouse model to the human protein in the same class.

Pfam, PIRSF, and PANTHER all associate GO classes (which include the class terms and definitions) with a homologous group of proteins or domains. We will use these associations to provide an initial set of relations between ProEvo domain and protein classes and GO classes.

Given the possibility of functional shift within a homologous group of proteins, we propose that the appropriate relationship between a ProEvo class and a GO class will be *has_ancestral_property*, meaning that all instances of the class descend from a common ancestor, and that, unless otherwise specified (see below), the properties of the ancestor are inherited by all instances of the class, i.e. by all descendants of the ancestor. However, a subset of proteins in a larger family might, under the influence of natural selection, diverge so greatly from their ancestors that they can be considered to form a class of their own. In such cases, the new class can *lose* attributes of its ancestor; that is, the attributes of the ancestral class are not conserved in all of its descendant classes, as has happened with the homologous proteins protein-glutamine gamma-glutamyltransferase and erythrocyte membrane protein band 4.2. We can handle such situations by introducing the relation *lacks_ancestral_property* to represent this process. Thus, the ancestral erythrocyte membrane protein band 4.2 *lacks_ancestral_property* of involvement in protein modification, but *has_ancestral_property* structural constituent of cytoskeleton.

ProForm connections

To support functional annotation and disease understanding, relations are defined between ProForm component classes and other appropriate ontologies and controlled vocabularies (Figure 1). Connection of protein forms to GO classes using appropriate relations will support accurate functional annotation. Relations defined between ProForm classes and the Disease Ontology (DO) will facilitate disease understanding. The Sequence Ontology (SO) will provide a structured controlled vocabulary to describe the consequences of gene mutations on the protein sequence. For attributes not yet defined in OBO Foundry ontologies, well-accepted controlled vocabularies will be adopted.

Using the parser described above, cross-references to other ontologies or knowledgebases (e.g., HUGO, GO, OMIM, RESID) were extracted from UniProtKB/Swiss-Prot and MGI entries and mapped to appropriate ontological/controlled vocabulary terms for selected human and mouse proteins of known disease phenotypes. GO annotations with experimental evidence were extracted from iProClass. The results were converted into annotations of relationships between PRO classes and those from other ontologies and were subjected to verification and literature-based curation. An example is given below.

A PRO example

Smad proteins are essential to serine/threonine kinase receptor signaling pathways regulated by phosphorylation. Smad2 undergoes phosphorylation at serines 465 and 467 upon activation of the transforming growth factor-beta (TGF-beta) type I receptor [37] (Figure 3). The phosphorylations permit association of Smad2 with Smad4, nuclear translocation of the complex, and regulation of transcription [38]. Therefore, the TGF-beta receptor-phosphorylated form is the active Smad2 form. Other forms of Smad2 are possible (Figure 4). We have curated a prototype PRO using Smad proteins from the human and mouse TGF-beta and bone morphogenetic signaling pathways [5]. Later versions will be expanded to additional proteins from these and other pathways.

Figure 5 illustrates the PRO structure for the Smad2 protein. Smad2 belongs to the "smad protein" family (source: PIRSF037286) and, more specifically, to the subfamily "receptor-regulated Smad protein, Smad2/Smad3 type" (source: PIRSF500455). Smad family proteins contain MH1 and MH2 domains (source: PF03165 and PF03166, respectively). The former is found in Smad-related proteins and nuclear factor 1 family proteins, whereas the MH2 domain is exclusively found in Smad proteins.

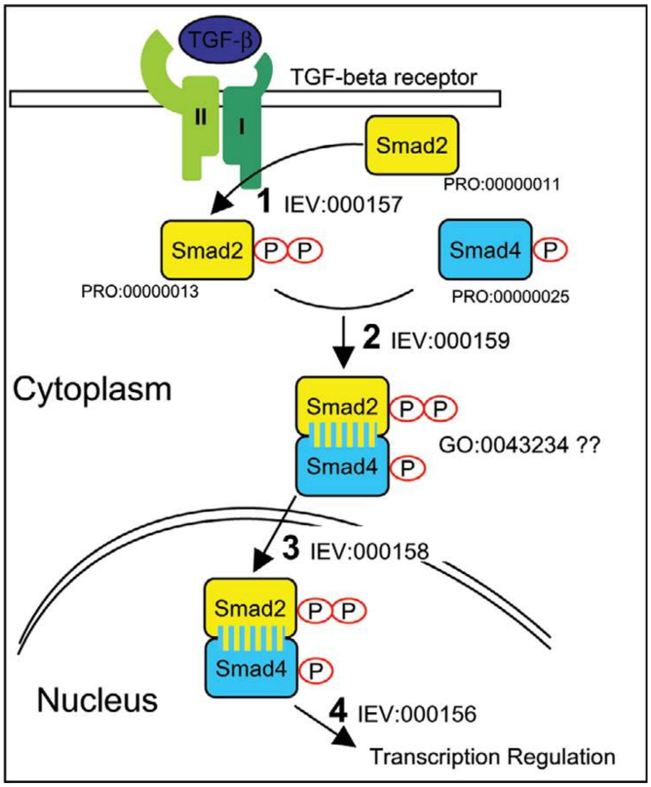


Figure 3
Smad2 component of the TGF-beta signaling pathway. Not all protein forms and pathway branches are indicated. The steps shown are preceded by phosphorylation of Smad4, TGF-beta binding to the receptor, and receptor phosphorylation. Step 1: Phosphorylation of Smad2 by TGF beta receptor I. Step 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.

Each gene may give rise to more than one PRO node, including a wild-type canonical protein plus any described splice and genetic variants. Relationships to GO, PSI-MOD and UMLS are listed under the corresponding node with the use of controlled vocabulary (information currently annotated using UMLS will eventually be replaced by DO). The terms for *has_function*, *has_modification*, *participates_in* and *located_in* are applied only to the appropriate forms based on experimental verification.

The active phosphorylated form of Smad2 (PRO:00000013) *located_in* nucleus, derives from Smad2 sequence 1 (PRO:00000011) (designated by the *derives_from* symbol ">" preceding the PRO accession number), which is *located_in* cytoplasm. Also, the phosphorylated form acquires the function-related terms "transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity," "Smad binding," and "transcription coactivator activity." Two other forms are derived from further phosphorylation of the active form, and represent the product obtained after regulation by other kinases. Phosphorylation by ERK-1 yields a form (PRO:00000014) with increased transcription coactivator activity [39] (there is currently no "modulation" ontology that provides this type of annotation), while phosphorylation by CAMK2 yields a form (PRO:00000015) that does not localize to the nucleus and thus is unable to coactivate transcription.

Smad2 has one splice form that lacks exon 3 (PRO:00000016). This form still maintains the characteristic functions of the TGF-beta receptor-activated form of Smad2, but can now bind directly to DNA (as can the closely-related Smad3 and other so-called R-Smads), and its transcription activity is further enhanced [40].

Finally, genetic variants related to disease are listed. Mutations in Smad2 have been found in colorectal carcinoma. TGF-beta signaling occurring during human colorectal carcinogenesis involves a shift in TGF-beta function, reducing the cytokine's anti-proliferative effect, while increasing actions that promote invasion and metastasis [41]. In the case of the variant with histidine-445 (PRO:00000019), signaling through the TGF-beta pathway is disrupted. The protein is expressed but is not phosphorylated.

Need for a protein ontology

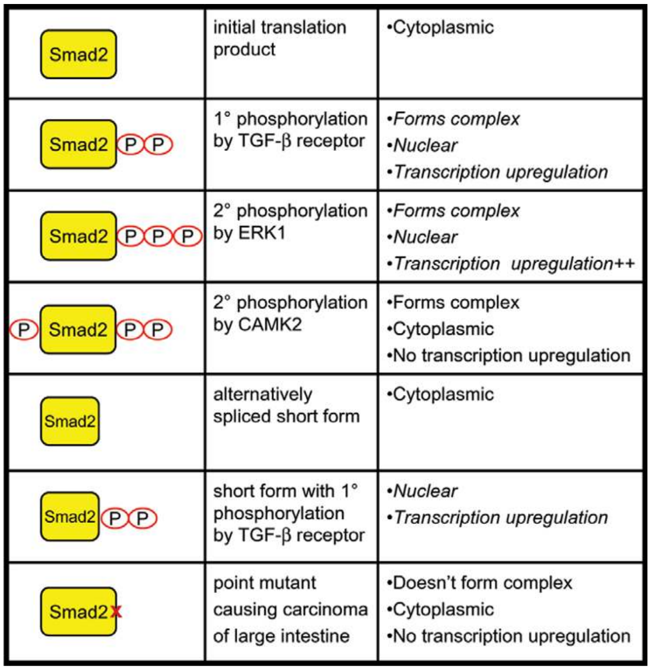
A protein ontology must fill two distinct needs: 1) a structure to support formal, computer-based inferences of shared attributes among homologous proteins; and 2) an explicit representation of the various forms of a given gene product.

Need for a protein evolution component

Protein sequence homology (i.e., descent from a common ancestral sequence) is the most widely used approach for annotating the putative functions of genes. While homology with a protein of known, experimentally characterized function is a critical tool for inferring the function of an uncharacterized protein, these inferences must be made carefully. Because there are no simple rules that can be applied consistently for all attributes of all proteins, homology-based inference methods can lead to errors. In a study of annotation consistency in the FlyBase, the primary cause (60%) for errors in GO annotations was determined to be incorrect homology-based inference [42]. However, all of the homology-based errors that were detected could be corrected using more rigorous whole-protein family/subfamily-based rules for functional inference, such as is done in the protein classification databases PANTHER and PIRSF. An ontology of protein evolution that explicitly models both whole proteins and parts of proteins (domains) will support formal, computer-based inferences of shared attributes among homologous proteins, and will enable more consistent, accurate and precise computational annotation. Furthermore, such an ontology will allow for a richer expression of relationships (including negatives such as "lacks" [25]) than is possible using conventional protein classification databases. This formalization will ensure rigorous application of experimental data to understanding protein-coding genes derived from high-throughput genome, cDNA, EST, or environmental sequencing projects. In addition, it will allow the transfer of described function/phenotypes of proteins from model organisms to human orthologs and may highlight potential candidates to explain a human disease (see below).

Need for a protein forms component

Alternative splicing and post-translational modifications give rise to the multiple products of a single gene, each of which can have different activities and different expression patterns. Nonetheless, the annotation information in most model organism and sequence databases is organized within a single entry, often without indicating which possible form is the correct object for annotation. Thus, annotation is associated with protein X when in fact it is specific to peptide Y derived from protein X, or to isoform Xa, or to a phosphorylated form of protein X; disease associations are more accurately ascribed to mutant forms of protein X. These various specific entities are natural components of pathway ontologies or databases such as INOH Event Ontology [17] or Reactome [19] (the latter does contain the relevant entities, but as accessions only that are not formed into an ontology structure). As biomedical data expand, it will be increasingly important to explicitly represent these protein forms so attributes are attached to the appropriate entity. For example, the GO



Smad2	initial translation product	•Cytoplasmic
Smad2 P P	1° phosphorylation by TGF-β receptor	•*Forms complex* •*Nuclear* •*Transcription upregulation*
Smad2 P P P	2° phosphorylation by ERK1	•*Forms complex* •*Nuclear* •*Transcription upregulation++*
P Smad2 P P	2° phosphorylation by CAMK2	•Forms complex •Cytoplasmic •No transcription upregulation
Smad2	alternatively spliced short form	•Cytoplasmic
Smad2 P P	short form with 1° phosphorylation by TGF-β receptor	•*Nuclear* •*Transcription upregulation*
Smad2✗	point mutant causing carcinoma of large intestine	•Doesn't form complex •Cytoplasmic •No transcription upregulation

Figure 4
Multiple possible Smad2 forms. Not all possibilities are indicated. The third column indicates the known properties for each form. Italicized text indicates those properties that are accurately reflected by the GO terms currently used to annotate human Smad2 in UniProt KB (SMAD2_HUMAN; accession Q15796).

terms that annotate the UniProtKB entry for human Smad2 indicate that the protein is a complex-forming nuclear transcription upregulator involved in the TGF-beta signaling cascade. While this does indeed describe the major function of Smad2, the annotation is only accurate for two out of the seven forms shown in Figure 4. The remaining forms either only partly fit this description or don't fit it at all.

PRO facilitates understanding of human disease

Mouse models can give valuable insight into human biology. As indicated in Figure 5, the human and mouse Smad2 have many common sequence forms. Alternative splicing of Smad2 exon 3 gives rise to a second distinct protein isoform. The phosphorylated short Smad2 isoform (PRO:00000018), unlike the full-length phosphorylated Smad2 (PRO:00000013), retains the direct DNA-binding activity (GO:0003677) common to every other receptor-regulated Smad (R-Smad; including Smads 1, 3, 5, and 8) [43]. Importantly, PRO shows that this form is common to mouse and human. Knockout mouse experiments indicate that Smad2 plays an essential role in patterning the embryonic axis and specification of definitive endoderm [44]. Mice that exclusively express the short isoform correctly specify the anterior-posterior axis and definitive endoderm and are viable and fertile, suggesting that the short form activates all essential target genes downstream of TGF-beta-related ligands early in development [45]. The direct comparison between specific mouse and human sequence forms facilitated by PRO can guide scientific inquiry. For example, experiments designed to elucidate the specific role of each human Smad2 isoform at different developmental stages are suggested by the information uncovered in the mouse counterparts. The activities of specific isoforms derived from variants that are agents of colorectal carcinomas could also be a factor to investigate.

Conclusion

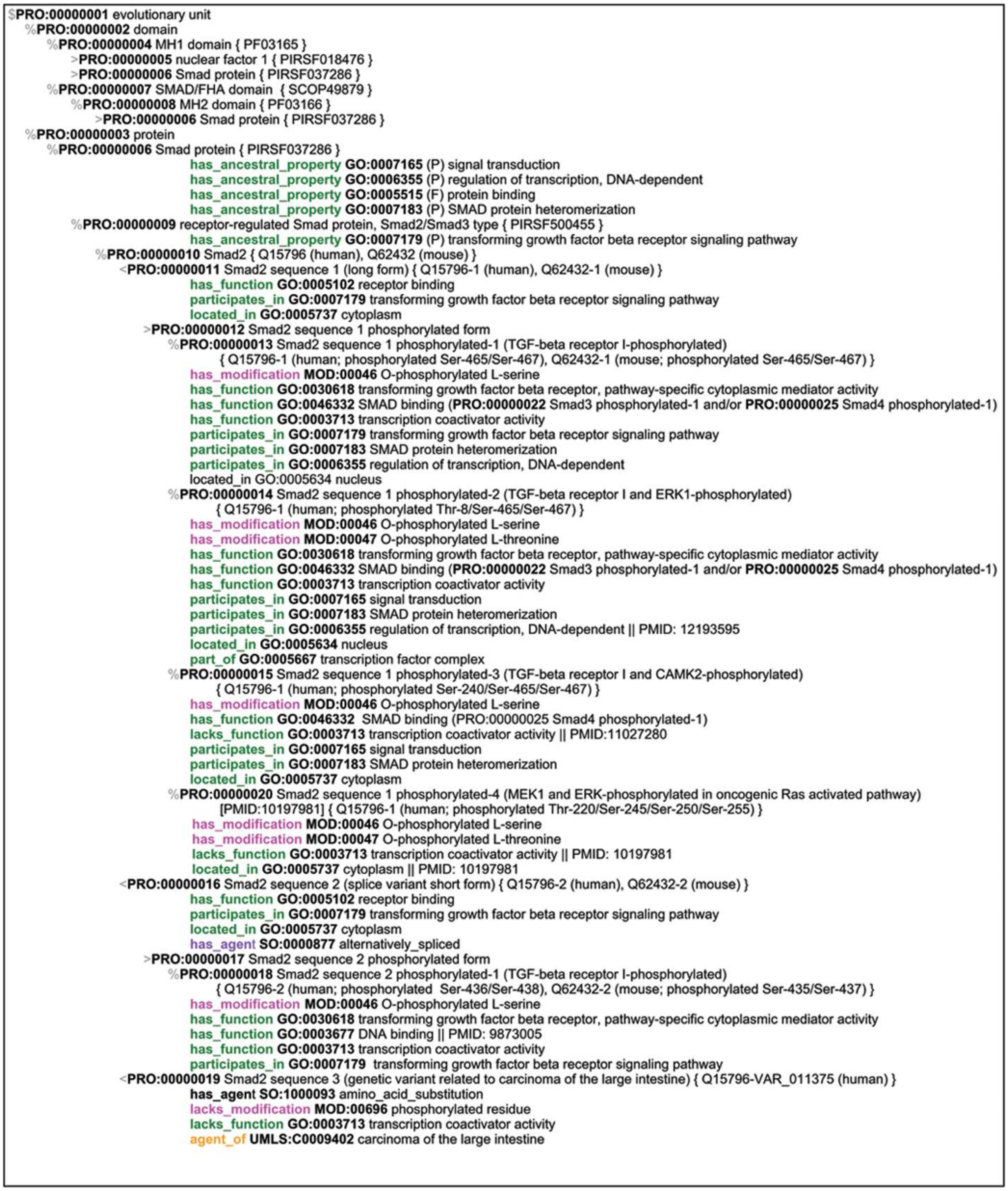
PRO is designed to be a formal, well-principled and extensible OBO Foundry ontology for proteins, with a basic set of well-defined relations to support semantic integration and machine reasoning. PRO development has begun by including classes relevant to human and mouse proteins in UniProtKB/Swiss-Prot and MGI, with a focus on disease-related proteins, but will be expanded on a system-by-system basis using pathways covered by the INOH pathway and Reactome databases. The framework described here can be adopted by scientists interested in curating PRO classes from other areas of interest.

The development of PRO is expected to create a cycle of improvement for both the ontology and the protein knowledgebases from which the initial information is extracted. For example, literature-based curation revealed that two of the modifications noted in the UniProtKB entry SMAD2_HUMAN occur in a single molecule (PRO:00000014). Such information can be fed back into the UniProtKB entry, along with the PRO node. Similar annotations throughout the database will, in turn, provide a richer information source for PRO.

PRO will have an impact beyond the knowledge contained therein. For example, essentially all homologous proteins in families referenced by PRO – irrespective of source organism – can be annotated using PRO terms, including the attributes from connected ontologies. In addition, by providing the means to reference specific protein forms, PRO will foster a precise interconnection between several ontologies. Comparison of information among related organisms and related ontologies is indispensable to human disease research.

List of abbreviations used

ICD10, International Classification of Diseases, Tenth Revision; SNOMED CT, Systemized Nomenclature of Medicine Clinical Terms; PSI-MOD, Proteomics Standards Initiative Protein Modification; PSI-MI, Proteomics Standards Initiative Molecular Interactions; INOH, Inte-



```
$PRO:00000001 evolutionary unit
  %PRO:00000002 domain
     %PRO:00000004 MH1 domain { PF03165 }
        >PRO:00000005 nuclear factor 1 { PIRSF018476 }
        >PRO:00000006 Smad protein { PIRSF037286 }
     %PRO:00000007 SMAD/FHA domain  { SCOP49879 }
        %PRO:00000008 MH2 domain { PF03166 }
           >PRO:00000006 Smad protein { PIRSF037286 }
  %PRO:00000003 protein
     %PRO:00000006 Smad protein { PIRSF037286 }
                        has_ancestral_property GO:0007165 (P) signal transduction
                        has_ancestral_property GO:0006355 (P) regulation of transcription, DNA-dependent
                        has_ancestral_property GO:0005515 (F) protein binding
                        has_ancestral_property GO:0007183 (P) SMAD protein heteromerization
        %PRO:00000009 receptor-regulated Smad protein, Smad2/Smad3 type { PIRSF500455 }
                        has_ancestral_property GO:0007179 (P) transforming growth factor beta receptor signaling pathway
           %PRO:00000010 Smad2 { Q15796 (human), Q62432 (mouse) }
              <PRO:00000011 Smad2 sequence 1 (long form) { Q15796-1 (human), Q62432-1 (mouse) }
                        has_function GO:0005102 receptor binding
                        participates_in GO:0007179 transforming growth factor beta receptor signaling pathway
                        located_in GO:0005737 cytoplasm
                 >PRO:00000012 Smad2 sequence 1 phosphorylated form
                    %PRO:00000013 Smad2 sequence 1 phosphorylated-1 (TGF-beta receptor I-phosphorylated)
                           { Q15796-1 (human; phosphorylated Ser-465/Ser-467), Q62432-1 (mouse; phosphorylated Ser-465/Ser-467) }
                        has_modification MOD:00046 O-phosphorylated L-serine
                        has_function GO:0030618 transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity
                        has_function GO:0046332 SMAD binding (PRO:00000022 Smad3 phosphorylated-1 and/or PRO:00000025 Smad4 phosphorylated-1)
                        has_function GO:0003713 transcription coactivator activity
                        participates_in GO:0007179 transforming growth factor beta receptor signaling pathway
                        participates_in GO:0007183 SMAD protein heteromerization
                        participates_in GO:0006355 regulation of transcription, DNA-dependent
                        located_in GO:0005634 nucleus
                    %PRO:00000014 Smad2 sequence 1 phosphorylated-2 (TGF-beta receptor I and ERK1-phosphorylated)
                           { Q15796-1 (human; phosphorylated Thr-8/Ser-465/Ser-467) }
                        has_modification MOD:00046 O-phosphorylated L-serine
                        has_modification MOD:00047 O-phosphorylated L-threonine
                        has_function GO:0030618 transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity
                        has_function GO:0046332 SMAD binding (PRO:00000022 Smad3 phosphorylated-1 and/or PRO:00000025 Smad4 phosphorylated-1)
                        has_function GO:0003713 transcription coactivator activity
                        participates_in GO:0007165 signal transduction
                        participates_in GO:0007183 SMAD protein heteromerization
                        participates_in GO:0006355 regulation of transcription, DNA-dependent || PMID: 12193595
                        located_in GO:0005634 nucleus
                        part_of GO:0005667 transcription factor complex
                    %PRO:00000015 Smad2 sequence 1 phosphorylated-3 (TGF-beta receptor I and CAMK2-phosphorylated)
                           { Q15796-1 (human; phosphorylated Ser-240/Ser-465/Ser-467) }
                        has_modification MOD:00046 O-phosphorylated L-serine
                        has_function GO:0046332  SMAD binding (PRO:00000025 Smad4 phosphorylated-1)
                        lacks_function GO:0003713 transcription coactivator activity || PMID:11027280
                        participates_in GO:0007165 signal transduction
                        participates_in GO:0007183 SMAD protein heteromerization
                        located_in GO:0005737 cytoplasm
                    %PRO:00000020 Smad2 sequence 1 phosphorylated-4 (MEK1 and ERK-phosphorylated in oncogenic Ras activated pathway)
                           [PMID:10197981] { Q15796-1 (human; phosphorylated Thr-220/Ser-245/Ser-250/Ser-255) }
                        has_modification MOD:00046 O-phosphorylated L-serine
                        has_modification MOD:00047 O-phosphorylated L-threonine
                        lacks_function GO:0003713 transcription coactivator activity || PMID: 10197981
                        located_in GO:0005737 cytoplasm || PMID: 10197981
              <PRO:00000016 Smad2 sequence 2 (splice variant short form) { Q15796-2 (human), Q62432-2 (mouse) }
                        has_function GO:0005102 receptor binding
                        participates_in GO:0007179 transforming growth factor beta receptor signaling pathway
                        located_in GO:0005737 cytoplasm
                        has_agent SO:0000877 alternatively_spliced
                 >PRO:00000017 Smad2 sequence 2 phosphorylated form
                    %PRO:00000018 Smad2 sequence 2 phosphorylated-1 (TGF-beta receptor I-phosphorylated)
                           { Q15796-2 (human; phosphorylated  Ser-436/Ser-438), Q62432-2 (mouse; phosphorylated Ser-435/Ser-437) }
                        has_modification MOD:00046 O-phosphorylated L-serine
                        has_function GO:0030618 transforming growth factor beta receptor, pathway-specific cytoplasmic mediator activity
                        has_function GO:0003677 DNA binding || PMID: 9873005
                        has_function GO:0003713 transcription coactivator activity
                        participates_in GO:0007179  transforming growth factor beta receptor signaling pathway
              <PRO:00000019 Smad2 sequence 3 (genetic variant related to carcinoma of the large intestine) { Q15796-VAR_011375 (human) }
                        has_agent SO:1000093 amino_acid_substitution
                        lacks_modification MOD:00696 phosphorylated residue
                        lacks_function GO:0003713 transcription coactivator activity
                        agent_of UMLS:C0009402 carcinoma of the large intestine
```

Figure 5
A PRO example using nodes and relationships illustrated by Smad2 protein. Not all possibilities are indicated. Cross-references to source of information (in curly braces) and description of sequence forms (in parentheses) are given for clarity. The symbols preceding each PRO accession are as follows: $: root; >: has_part (for domains) or derives_from (for proteins); <: variant_of.

grating Network Objects with Hierarchies; PANTHER, Protein ANalysis THrough Evolutionary Relationships.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DAN, CNA, WCB, ZH, HL, BS, CHW participated in the production of the manuscript. HL performed the data mining to produce an automatically-generated "pre-PRO." CNA curated the Protein Ontology for Smad2 (and related proteins of the TGF-B signaling pathway). TCC and CNA curated the PIRSF protein families for Smad2 (and related proteins of the TGF-B signaling pathway). DAN, CNA, WCB, JB, ZH, BS, and CHW formulated the Protein Ontology framework. All authors read and approved the final manuscript.

Acknowledgements

Paul Thomas (SRI) provided critical input on the ProEvo component. Funding for PRO development is provided by NIH grant 1 R01 GM080646-01. Support for B. Smith's contribution was provided by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 U 54 HG004028. Support for J. Blake's contribution was provided by NIH P41 grants HG 002273 and HG 003300.

This article has been published as part of *BMC Bioinformatics* Volume 8 Supplement 9, 2007: First International Workshop on Text Mining in Bioinformatics (TMBio) 2006. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/8?issue=S9.

References

1. Bard J: **Ontologies: Formalising biological knowledge for bioinformatics.** *Bioessays* 2003, **25:**501-506.
2. Blake JA: **Bio-ontologies – fast and furious.** *Nat Biotechnol* 2004, **22:**773-774.
3. Blake JA, Bult CJ: **Beyond the data deluge: data integration and bio-ontologies.** *J Biomed Inform* 2006, **39:**314-320.
4. **Open Biomedical Ontologies** [http://obo.sourceforge.net]
5. **Protein Ontology** [http://pir.georgetown.edu/pro]
6. **The OBO Foundry** [http://obofoundry.org/]
7. Gene Ontology Consortium: **The Gene Ontology (GO) project in 2006.** *Nucleic Acids Res* 2006, **34:**D322-326.
8. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biol* 2005, **6:**R44.
9. **Disease Ontology** [http://diseaseontology.sourceforge.net/]
10. **International Classification of Diseases (ICD)** [http://www.who.int/classifications/icd/en/]
11. **SNOMED Clinical Terms** [http://www.snomed.org/]
12. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004, **32:**D267-270.
13. **Proteomics Standards Initiative Protein Modification** [http://psidev.sourceforge.net/mod/]
14. Garavelli JS: **The RESID Database of Protein Modifications as a resource and annotation tool.** *Proteomics* 2004, **4:**1527-1533.
15. UniProt Consortium: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007, **35:**D193-197.
16. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, Roechert B, Poux S, Jung E, Mersch H, Kersey P, Lappe M, Li Y, Zeng R, Rana D, Nikolski M, Husi H, Brun C, Shanker K, Grant SG, Sander C, Bork P, Zhu W, Pandey A, Brazma A, Jacq B, Vidal M, Sherman D, Legrain P, Cesareni G, Xenarios I, Eisenberg D, Steipe B, Hogue C, Apweiler R: **The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data.** *Nat Biotechnol* 2004, **22:**177-183.
17. Kushida T, Takagi T, Fukuda KI: **Event Ontology: A pathway-centric ontology for biological processes.** *Pac Symp Biocomput* 2006, **11:**152-163.
18. Yamamoto S, Asanuma T, Takagi T, Fukuda KI: **The molecule role ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature.** *Comparative and Functional Genomics* 2004, **5:**528-536.
19. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, Lewis S, Birney E, Stein L: **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Res* 2005, **33:**D428-432.
20. Sidhu AS, Dillon TS, Chang E, *et al.*: **Protein Ontology: Data Integration using Protein Ontology.** In *Database Modeling in Biology: Practices and Challenges* Edited by: *Ma Z, Chen JY. New York, Springer Inc*; 2006:39-60.
21. Sahoo SS, Thomas C, Sheth A, York WS, Tartir S: **Knowledge modeling and its application in life sciences: a tale of two ontologies.** *Proceedings of the 15th International Conference on World Wide Web* 2006:317-326 [http://portal.acm.org/citation.cfm?id=1135826&jmp=abstract&coll=GUIDE&dl=GUI]. *New York, ACM Press*
22. Smith B, Ceuster W, Klagger B, Kohler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C: **Relations in Biomedical Ontologies.** *Genome Biol* 2005, **6:**R46.
23. Burglin TR, Cassata G: **Loss and gain of domains during evolution of cut superclass homeobox genes.** *Int J Dev Biol* 2002, **46:**115-123.
24. Shimeld SM: **Characterization of AmphiF-spondin reveals the modular evolution of chordate F-spondin genes.** *Mol Biol Evol* 1998, **15:**1218-1223.
25. Ceusters W, Elkin P, Smith B: **Referent tracking: The problem of negative findings.** *Stud Health Technol Inform* 2006, **124:**741-746.
26. Mandal D, Moitra PK, Basu J: **Mapping of a spectrin-binding domain of human erythrocyte membrane protein 4.2.** *Biochem J* 2002, **364:**841-847.
27. Polakowska RR, Eickbush T, Falciano V, Razvi F, Goldsmith LA: **Organization and evolution of the human epidermal keratinocyte transglutaminase I gene.** *Proc Natl Acad Sci USA* 1992, **89:**4476-4480.
28. Korsgren C, Lawler J, Lambert S, Speicher D, Cohen CM: **Complete amino acid sequence and homologies of human erythrocyte membrane protein band 4.2.** *Proc Natl Acad Sci USA* 1990, **87:**613-617.
29. Boeshans KM, Mueser TC, Ahvazi B: **A three-dimensional model of the human transglutaminase 1: insights into the understanding of lamellar ichthyosis.** *J Mol Model* 2007, **13:**233-246.
30. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Res* 2004, **32:**D226-229.
31. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, Eddy SR, Sonnhammer EL, Bateman A: **Pfam: clans, web tools and services.** *Nucleic Acids Res* 2006, **34:**D247-251.
32. Wu CH, Nikolskaya A, Huang H, Yeh L-S, Natale DA, Vinayaka CR, Hu Z, Mazumder R, Kumar S, Kourtesis P, Ledley RS, Suzek BE, Arminski L, Chen Y, Zhang J, Cardenas JL, Chung S, Castro-Alvear J, Dinkov G, Barker WC: **PIRSF family classification system at the Protein Information Resource.** *Nucleic Acids Res* 2004, **32:**D112-114.
33. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD: **The PANTHER database of protein families, subfamilies, functions and pathways.** *Nucleic Acids Res* 2005, **33:**D284-288.
34. Bork P, Sander C, Valencia A: **Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, and galactokinase families of sugar kinases.** *Protein Sci* 1993, **2:**31-40.
35. Blake JA, Eppig JT, Bult CJ, Kadin JA, Richardson JE: **Mouse Genome Database Group. The Mouse Genome Database (MGD): updates and enhancements.** *Nucleic Acids Res* 2006, **34:**D562-567.

36. Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC: **The iProClass integrated database for protein functional analysis.** *Comput Biol Chem* 2004, **28:**87-96.
37. Kretzschmar M, Liu F, Hata A, Doody J, Massague J: **The TGF-beta family mediator Smad1 is phosphorylated directly and activated functionally by the BMP receptor kinase.** *Genes Dev* 1997, **11:**984-995.
38. Abdollah S, Macias-Silva M, Tsukazaki T, Hayashi H, Attisano L, Wrana JL: **TbetaRI phosphorylation of Smad2 on Ser465 and Ser467 is required for Smad2–Smad4 complex formation and signaling.** *J Biol Chem* 1997, **272:**27678-27685.
39. Funaba M, Zimmerman CM, Mathews LS: **Modulation of Smad2-mediated signaling by extracellular signal-regulated kinase.** *J Biol Chem* 2002, **277:**41361-41368.
40. Yagi K, Goto D, Hamamoto T, Takenoshita S, Kato M, Miyazono K: **Alternatively spliced variant of Smad2 lacking exon 3. Comparison with wild-type Smad2 and Smad3.** *J Biol Chem* 1999, **274:**703-709.
41. Matsuzaki K: **Smad3 phosphoisoform-mediated signaling during sporadic human colorectal carcinogenesis.** *Histol Histopathol* 2006, **21:**645-662.
42. Mi H, Vandergriff J, Campbell M, Narechania A, Majoros W, Lewis S, Thomas PD, Ashburner M: **Assessment of genome-wide protein function classification for Drosophila melanogaster.** *Genome Res* 2003, **13:**2118-2128.
43. Massague J, Seoane J, Wotton D: **Smad transcription factors.** *Genes Dev* 2005, **19:**2783-2810.
44. Weinstein M, Yang X, Deng C: **Functions of mammalian Smad genes as revealed by targeted gene disruption in mice.** *Cytokine Growth Factor Rev* 2000, **11:**49-58.
45. Dunn NR, Koonce CH, Anderson DC, Islam A, Bikoff EK, Robertson EJ: **Mice exclusively expressing the short isoform of Smad2 develop normally and are viable and fertile.** *Genes Dev* 2005, **19:**152-163.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral