Framework for a Protein Ontology

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

[1]Department of Biochemistry and Molecular & Cellular Biology
Georgetown University Medical Center
3300 Whitehaven St., NW
Washington, DC 20007, USA
{dan5,cna5,wb8,tc245,zh9,wuc}
@georgetown.edu

[3]Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University
Room 176, Building D
Washington, DC 20057, USA
hl224@georgetown.edu

[2]The Jackson Laboratory
600 Main St.
Bar Harbor, ME 04609, USA
jblake@informatics.jax.org

[4]Department of Philosophy
State University of New York at Buffalo
Park Hall
Buffalo, NY 12460, USA
phismith@buffalo.edu

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 exist that describe the properties that can be attributed to proteins; for example, protein functions are described by Gene Ontology, while human diseases are described by Disease Ontology. There is, however, a gap in the current set of ontologies—one that describes the protein entities themselves and their relationships. We have designed a 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 modification). PRO will allow the specification of relationships between PRO, GO and other OBO Foundry ontologies. Here we describe the initial development of PRO, illustrated using human proteins from the TGF-beta signaling pathway (http://pir.georgetown.edu/pro).

Categories and Subject Descriptors

E.2 [Data Storage Representations]: *Object representation;* H.2.8 [Database Management]: Database Applications - *Data mining, Scientific databases;* I.2.4 [Artificial Intelligence]: Knowledge Representation Formalisms and Methods - *Relation systems;* J.3 [Life and Medical Sciences]: *Biology and genetics.*

General Terms

Documentation, Reliability, Standardization

TMBIO'06, November 10, 2006, Arlington, Virginia, USA.

Keywords

Ontology, Protein, OBO, OBO Foundry, PIRSF, PANTHER, SCOP, Pfam, GO

1. INTRODUCTION

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 [3][4][5]. The Open Biomedical Ontologies (OBO) (http://obo.sourceforge.net) 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. Here we describe the initial development of a **PR**otein **O**ntology (**PRO**) to describe the relationships of proteins and protein evolutionary families (ontology for protein evolution), to delineate the multiple protein forms of a gene locus (ontology for protein forms), and to interconnect existing ontologies.

1.1 Biomedical Ontologies

For an ontology to be of public 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.

The OBO Foundry (http://obofoundry.org/) has outlined a set of principles specifying best practices in ontology development to foster interoperability of ontologies and ensure a gradual improvement of quality and formal rigor. Ontologies in the OBO Foundry are required to be well-documented, adopt a common formal language, and be developed in a collaborative manner. The following summarizes several candidate OBO Foundry ontologies that are related to PRO.

1.1.1 Gene Ontology (GO)

The Gene Ontology (GO) is by far the most widely used ontology in any discipline [16]. It aims to formalize the capture and representation of information about biological processes, molecular functions, and cellular components through three mutually independent hierarchies. GO has been used to annotate the genes of humans and a variety of model organisms, and has facilitated both in-depth understanding of biology within a single organism and comparing biological processes across multiple organisms.

1.1.2 Sequence Ontology (SO)

The Sequence Ontology (SO) provides a rich set of terms, relations, and definitions for genome and chromosome annotation [11]. A subset of SO terms addresses the consequences of gene mutation on protein products; for example, whether the mutation decreases or eliminates protein activity. Such terms are relevant to protein entities, and thus will be connected to the corresponding PRO entries.

1.1.3 Disease Ontology (DO)

DO (http://diseaseontology.sourceforge.net/) is a controlled medical vocabulary to facilitate the mapping of diseases and associated conditions to medical coding systems such as ICD9CM and SNOMED. In addition, DO provides cross-references to the Unified Medical Language System (UMLS) [7].

1.1.4 Other Protein-Related Ontologies

The PSI-MOD ontology (http://psidev.sourceforge.net/mod/) 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 [15] terms, a controlled vocabulary for defining modification features of protein entries in the UniProt Knowledgebase (UniProtKB) [12]. PSI-MI [17] and INOH Event Ontology (EO) [22] are ontologies that, in part, describe the events of protein interaction. Finally, the Molecule Role Ontology [38], another ontology developed for INOH pathway (http://www.inoh.org/) curation, contains molecular functional group names, abstract molecule names and concrete molecule names manually collected from literature. The structure of each of these resources aligns well with PRO.

Two other ontologies are designed for database integration or annotation. Protein Ontology (PO) [31] includes terms and relationships to describe attributes of individual protein forms, but does not include the protein forms themselves. ProPreO [29] is an ontology that enables a detailed description of proteomics experimental processes and data.

2. PROTEIN ONTOLOGY DEVELOPMENT

The development of PRO will proceed by taking a minimalist approach to defining relations and connections to other ontologies, and by taking a pragmatic approach to populating the ontology with terms and their annotations. We make use of existing relationships, such as those defined in the Relations Ontology [32], whenever possible. Where needed, well-defined relationships will be created according to the guidelines of the OBO Foundry. For annotation of PRO classes, connections to other ontologies will be used. Terms and annotations will be included by mining various sources of information using computational methods. This initial set of PRO records will be manually adjusted.

An overview of PRO is provided in Figure 1. Note that, while there will not be orthogonal ontologies within PRO, it is natural to view it as having two main components: a protein evolution component (ProEvo), and a protein forms component (ProForm)

2.1 Protein Evolution Component (ProEvo)

In order to reason over evolutionary relationships between proteins using an ontology-based formalism, we defined classes that represent protein sets (*i.e.*, families or clusters of related proteins). The classes are designed to reflect what we know about proteins and their evolution: (1) the tree-like structure of protein evolution, (2) the existence of two different heritable units: what are called “domains” (parts of proteins), and whole proteins (which can be made up of combinations of domains).

The diversity of proteins we find today in living organisms can be grouped into protein families, each member of which 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. It is 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 units of protein-coding sequence—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). 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., [8][30]). 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 [9].

Note that the relationships between ProEvo classes will not be based on function, as for the GO molecular function ontology, but on evolutionary relatedness. In many cases the distinction will not be obvious. However, consider the case of erythrocyte membrane protein band 4.2 (EPB4.2), a major component of the red blood cell membrane skeleton [23] that was co-opted from an ancestral class of protein-glutamine gamma-glutamyltransferases [28], and subsequently has lost that ancestral function [20]. In the GO molecular function ontology, the appropriate parent term for EPB4.2 would be “constituent of cytoskeleton” (GO:0005200). For PRO, the parent would be “protein-glutamine gamma-glutamyltransferase.”

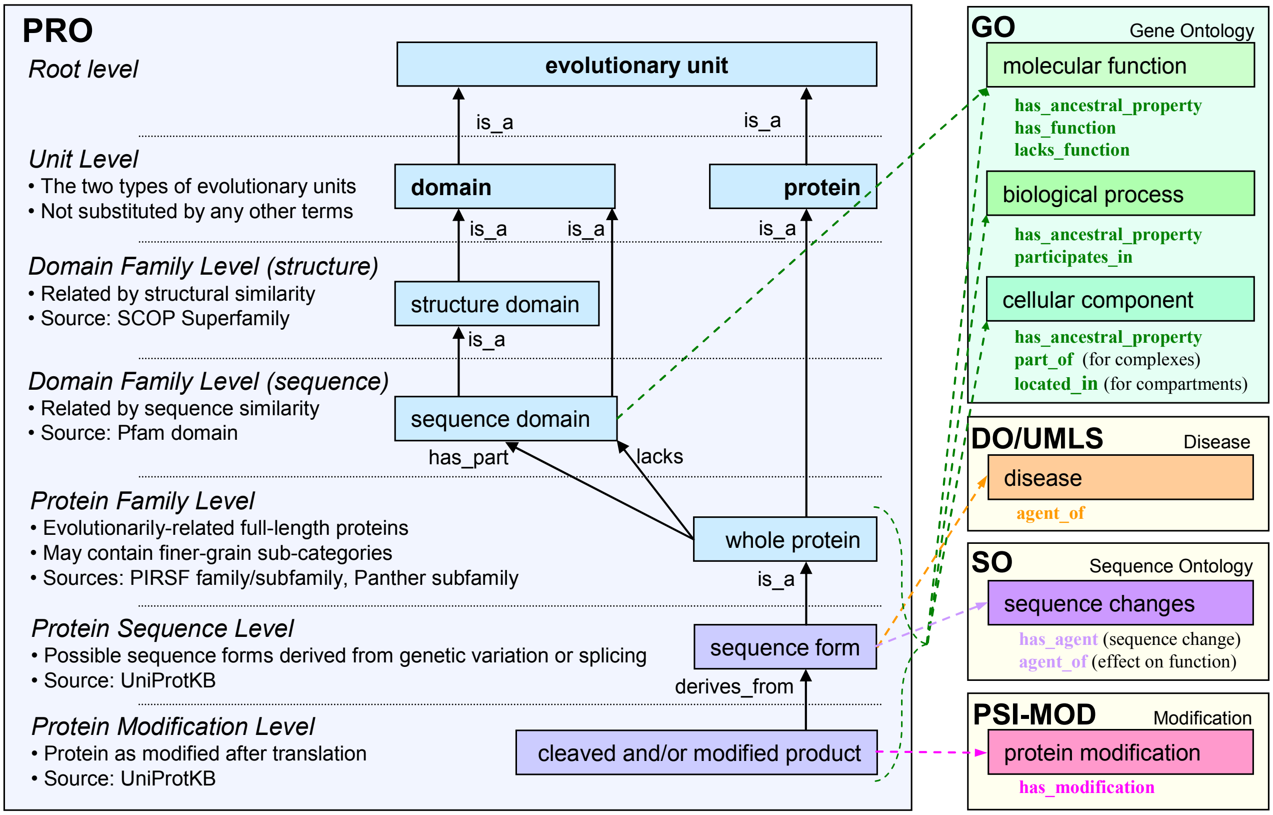
PRO
Root level
evolutionary unit
is_a
is_a
Unit Level
• The two types of evolutionary units
• Not substituted by any other terms
domain
protein
is_a
is_a
is_a
Domain Family Level (structure)
• Related by structural similarity
• Source: SCOP Superfamily
structure domain
is_a
Domain Family Level (sequence)
• Related by sequence similarity
• Source: Pfam domain
sequence domain
has_part
lacks
Protein Family Level
• Evolutionary-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.

2.1.1 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 the task 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 each approach is most appropriate for, from the most distant to the closest.

2.1.1.1 Structure-Based Clusters of "Remote" Domain Homology: SCOP

SCOP (Structural Classification of Proteins) [2] is arranged hierarchically into four levels: *class*, *fold*, *superfamily* and *family*. Homology can be asserted for proteins in the same family based on sequence data alone and for proteins in the same superfamily based on 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, for the purposes of PRO, only the SCOP superfamily and family data are suitable.

2.1.1.2 Sequence-Based Clusters of "Close" Domain Homology: Pfam

Pfam domain families [13] 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.

2.1.1.3 Clusters of Protein Homology: PIRSF

The PIRSF family classification system provides protein classification from superfamily to subfamily levels in a network structure to reflect evolutionary relationship of full-length proteins and domains [36]. 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 full-length proteins allows annotation of biological functions, biochemical activities, and sequence features that are family specific, while an understanding of the domain architecture of a protein provides insight into general functional and structural properties as well as into complex evolutionary mechanisms.

2.1.1.4 Functionally-Diverged Subfamilies: PANTHER

A PANTHER subfamily [26] 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.

2.1.2 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. This information is readily accessible via the ID mapping service based on iProClass at PIR. 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 [34]. Thus, each PRO class will have an incremented number as its accession (e.g., PRO:00000001) and a source-database identifier as its ID (e.g., PRO:PIRSF0000001).

2.2 Protein Forms Component (ProForm)

A number of different protein entities 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 are involved 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. 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, this is just a more specific way of asserting that "new entity *created_from* old entity." We will examine pre-existing ontologies to establish relationships suitable for our use. Primarily, we will use OBO's Relations Ontology as a source of well-defined relationships, and expand to relationships from other ontologies as needed. Thus, instead of *modified_from* in the example above, we say "new entity *derives_from* old entity;" the two relations are identical, and the latter is already part of the core set of relations [32]. However, this relation does not accurately describe two variations of the same gene product, nor does the *is_a* relation. Therefore, we use a new relation *variant_of* for this situation.

2.2.1 Populating ProForm Classes: Resources

Both the richness and primary use of an ontology stem from the diversity and comprehensiveness of its classes, and we intend to capture the diverse forms that a protein can take. UniProtKB/Swiss-Prot contains information on mutation, splice variants, protein cleavage, and post-translational modification. These data, found within the controlled vocabulary of FT (feature) lines or free text of CC (comment) lines, will be used to populate the appropriate classes. Other sources of data include MGI [6] and iProClass [35].

2.2.2 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 any annotation labeled as "by similarity," "potential," or "probable." There are three kinds of entities considered by the parser: isoforms, variants, and cleavage and modification products. Cross-references to other ontologies or knowledgebases were also extracted. GO annotations with the Traceable Author Statement evidence code were extracted from iProClass.

3. CONNECTING OTHER ONTOLOGIES

Several ontologies—notably, GO and DO—are pertinent to protein annotation, but are unable to connect directly to proteins 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 Z.

3.1 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 each associate GO terms 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. The associations for domain classes will be restricted to GO molecular function and will not be curated. Connections to other external ontologies will not initially be attempted.

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 a larger class might, under the influence of natural selection, diverge so greatly from its ancestors that it can be considered a class of its 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.

3.2 ProForm Connections

To support functional annotation and disease understanding, relations will be defined between ProForm component classes and other appropriate ontologies and controlled vocabularies (Figure 1). Connection of protein forms to GO terms 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.

ProForm connections will be curated by extracting existing annotations from the UniProtKB/Swiss-Prot and MGI entries and mapping them to appropriate ontological/controlled vocabulary terms for selected human and mouse proteins of known disease phenotypes. The parser described above extracts text in UniProtKB and converts the results into annotations of relationships to other ontologies. An example is given below.

4. A PRO EXAMPLE

Smad proteins are essential to serine/threonine kinase receptor signaling pathways regulated by phosphorylation. Smad 2 undergoes phosphorylation at serines 465 and 467 upon activation of the transforming growth factor-beta (TGF-beta) type I receptor [21] (Figure 2). The phosphorylations permit association with Smad 4, nuclear translocation and regulation of transcription [1]. Therefore, the receptor-phosphorylated form is the active entity. We have curated a prototype PRO using proteins from the TGF-beta signaling pathway (http://pir.georgetown.edu/pro). Later versions will include other pathways.

Figure 3 illustrates the PRO structure for the Smad 2 protein. Smad 2 is a whole protein of the "smad protein" family (source: PIRSF037286) and, more specifically, to the subfamily "receptor-regulated Smad protein, Smad 2/Smad 3 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.

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 object 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 Smad 2 (PRO:00000013), *located_in* nucleus, derives from Smad 2 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 entities are derived from further phosphorylation of the active form, and represent the product obtained after regulation by other kinases. ERK-1 phosphorylation (PRO:00000014) increases the transcription co-activation activity [14] (there is currently no "modulation" ontology that provides this type of annotation). CAMK2 phosphorylation (PRO:00000015) prevents nuclear localization and thus inhibits its transcription co-activation function.

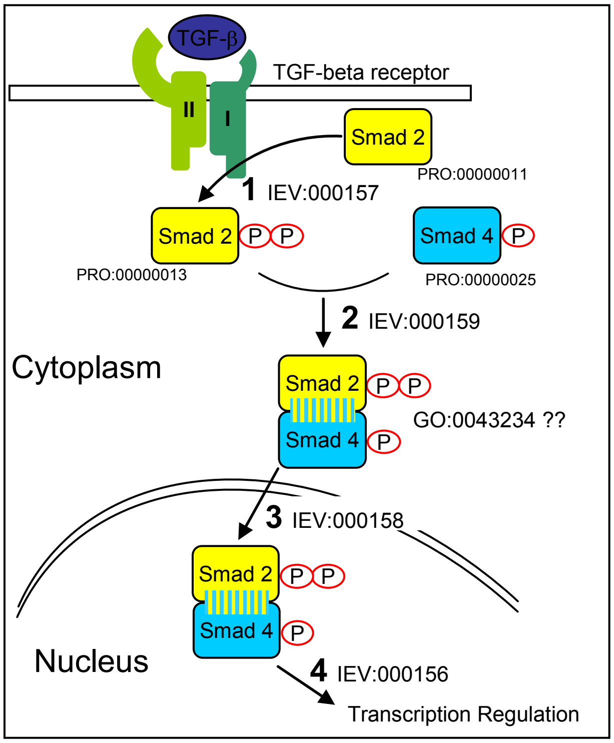

Figure 2. Smad 2 component of the TGF-beta signaling pathway. The steps shown are preceded by phosphorylation of Smad 4, TGF-beta binding to the receptor, and receptor phosphorylation. Step 1: Phosphorylation of Smad 2 by TGF beta receptor I. Step 2: Complex formation of R-smad and Smad 4. Step 3: Nuclear import of R-smad:Smad 4. Step 4: Binding of R-smad:Smad 4 complex coactivator to responsive element.

Smad 2 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 Smad 2, but can now bind directly to DNA (as can the closely-related Smad 3 and other so-called R-Smads), and its transcription activity is further enhanced [37].

Finally, genetic variants related to disease are listed. Mutations in Smad 2 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 [25]. 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.

5. 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.

5.1 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 critical tool for inferring the function of an uncharacterized

```
$PRO:00000001 evolutionary unit[a]
    %PRO:00000002 domain
        %PRO:00000004 MH1 domain { PF03165 }[b]
            >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, Smad 2/Smad 3 type { PIRSF500455 }[c]
                                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 protein 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, TGF-beta receptor I-phosphorylated { Q15796-1-P1 (human), Q62432-1-P1 (mouse) }
                                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 Smad 4 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, TGF-beta receptor I and ERK1-phosphorylated { Q15796-1-P2 (human) }
                                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 Smad 4 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, TGF-beta receptor I and CAMK2-phosphorylated { Q15796-1-P3 (human) }
                                has_modification MOD:00046 O-phosphorylated L-serine
                                has_function GO:0046332  SMAD binding (PRO:00000025 Smad 4 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:00000016 Smad2 sequence 2 (short form) - splice variant { Q15796-2 (human), Q62432-2 (mouse) }
                                has_agent SO:0000877 alternatively_spliced
                        >PRO:00000017 Smad2 sequence 2 phosphorylated form
                            %PRO:00000018 Smad2 sequence 2, TGF-beta receptor I-phosphorylated { Q15796-2-P1 (human), Q62432-2-P1 (mouse) }
                                has_modification MOD:00696 phosphorylated residue
                                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
```

[a] The symbols preceding each PRO accession are as follows:

$	root	>	has_part	(for domains) <	variant_of
%	is_a	>	derives_from	(for proteins)	

[b] Text in curly braces indicates the PRO ID, typically derived from the source of the class

[c] Not all examples shown.

Figure 3. A PRO example (nodes and relationships illustrated by Smad 2 protein)

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. The largest single reason for errors in FlyBase GO annotations was incorrect homology-based inference, accounting for 60% of the total number of errors [27]. 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. 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).

5.2 Need for a Protein Forms Component

The multiple products of a single gene can have different activities and expression patterns. Nonetheless, the annotation information in most model organism and sequence databases is organized within a single entry, often without indicating which of the specific forms are the correct objects 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.

5.3 PRO Facilitates Understanding of Human Disease

Mouse models can give valuable insight into human biology. As indicated in Figure 3, the human and mouse Smad 2 have many common sequence forms. Alternative splicing of Smad 2 exon 3 gives rise to a second distinct protein isoform. The phosphorylated short Smad 2 isoform (PRO:10000497), unlike the full-length phosphorylated Smad 2 (PRO:10000493), retains the direct DNA-binding activity (GO:0003677) common to every other regulatory Smad (R-Smad; including Smads 1, 3, 5, and 8) [24]. Importantly, PRO shows that this form is common to mouse and human. Knockout mouse experiments indicate that Smad 2 plays an essential role in patterning the embryonic axis and specification of definitive endoderm [33]. 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 [10]. The direct comparison between specific mouse and human sequence forms facilitated by PRO can lead to scientific discovery. That is, the information uncovered in mice can be used to look into the human counterparts. For example, experiments designed to elucidate the specific role of the human Smad 2 isoform at specific developmental stages are suggested. Also, the different activities of these isoforms, as revealed both in human and mouse (short isoform binds DNA with increased transcription cofactor activity), could be a factor to investigate in those variants that are agents of colorectal carcinomas.

6. CONCLUDING REMARKS

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 will initially include only human and mouse proteins in UniProtKB/Swiss-Prot and MGI, with a focus on disease-related proteins. PRO will be built on a system-by-system basis using pathways covered by the INOH pathway and Reactome [19] databases.

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 PRO families—irrespective of source organism—can be annotated using PRO terms, including the attributes from connected ontologies. Comparison of information among related organisms and related ontologies is indispensable to human disease research.

7. ACKNOWLEDGMENTS

Paul Thomas (SRI) provided critical input on the ProEvo component. Support for B. Smith's contribution was provided in part by the National Institutes of Health through the NIH Roadmap for Medical Research, Grant 1 U 54 HG004028.

8. REFERENCES

[1] Abdollah, S., Macias-Silva, M., Tsukazaki, T., Hayashi, H., Attisano, L., and Wrana, J.L. TbetaRI phosphorylation of Smad2 on Ser465 and Ser467 is required for Smad2-Smad4 complex formation and signaling. *J Biol Chem.* 272:27678-85, 1997.

[2] Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*. 32:D226-9, 2004.

[3] Bard, J. Ontologies: Formalising biological knowledge for bioinformatics. *Bioessays.* 25(5):501-506, 2003.

[4] Blake, J.A. Bio-ontologies--fast and furious. *Nat Biotechnol.* 22:773-774, 2004.

[5] Blake, J.A., and Bult, C.J. Beyond the data deluge: data integration and bio-ontologies. *J Biomed Inform.* 39: 314-320, 2006.

[6] Blake, J.A., Eppig, J.T., Bult, C.J., Kadin, J.A., and Richardson, J.E.; Mouse Genome Database Group. The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res.* 34:D562-7, 2006.

[7] Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32:D267-70, 2004.

[8] Burglin, T.R., and Cassata, G. Loss and gain of domains during evolution of cut superclass homeobox genes. *Int J Dev Biol.* 46:115-23, 2002.

[9] Ceusters, W., Elkin, P., and Smith, B. Referent tracking: The problem of negative findings. *Stud Health Technol Inform.*, in press, 2006.

[10] Dunn, N.R., Koonce, C.H., Anderson, D.C., Islam, A., Bikoff, E.K., and Robertson, E.J. Mice exclusively expressing the short isoform of Smad2 develop normally and are viable and fertile. *Genes Dev.* 19:152-63, 2005.

[11] Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* 6:R44, 2005.

[12] Farriol-Mathis, N., Garavelli, J.S., Boeckmann, B., Duvaud,

S., Gasteiger, E., Gateau, A., Veuthey, A.L., and Bairoch, A. Annotation of post-translational modifications in the Swiss-Prot knowledge base. *Proteomics* 4:1537-50, 2004.

[13] Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L., and Bateman, A. Pfam: clans, web tools and services. *Nucleic Acids Res*. 34:D247-51, 2006.

[14] Funaba, M., Zimmerman, C.M., and Mathews, L.S. Modulation of Smad2-mediated signaling by extracellular signal-regulated kinase. *J Biol Chem.* 277:41361-8, 2002.

[15] Garavelli, J.S. The RESID Database of Protein Modifications as a resource and annotation tool. *Proteomics* 4:1527-33, 2004.

[16] Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* 34:D322-6, 2006.

[17] 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, S.G., 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., and Apweiler, R. The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data. *Nat Biotechnol.* 22:177-83, 2004.

[19] Joshi-Tope, G., Gillespi,e M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, GR., Wu, G.R., Matthews, L., Lewis, S., Birney, E., and Stein, L. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. 33:D428-32, 2005.

[20] Korsgren, C., Lawler, J., Lambert, S., Speicher, D., and Cohen, C.M. Complete amino acid sequence and homologies of human erythrocyte membrane protein band 4.2. *Proc Natl Acad Sci U S A*. 87:613-617, 1990.

[21] Kretzschmar, M., Liu, F., Hata, A., Doody, J., and Massague, J. The TGF-beta family mediator Smad1 is phosphorylated directly and activated functionally by the BMP receptor kinase. *Genes Dev.* 11:984-95, 1997.

[22] Kushida, T., Takagi, T., and Fukuda, K.I. Event Ontology: A pathway-centric ontology for biological processes. *Pacific Symposium on Biocomputing* 11:152-163, 2006.

[23] Mandal, D., Moitra, P.K., and Basu, J. Mapping of a spectrin-binding domain of human erythrocyte membrane protein 4.2. *Biochem J.* 364:841-7, 2002.

[24] Massague, J., Seoane, J., and Wotton, D. Smad transcription factors. *Genes Dev*. 19:2783-810, 2005.

[25] Matsuzaki K.. Smad3 phosphoisoform-mediated signaling during sporadic human colorectal carcinogenesis. *Histol Histopathol*. 21:645-62, 2006.

[26] Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M.J., Kitano, H., and Thomas, P.D. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* 33:D284-8, 2005.

[27] Mi, H., Vandergriff, J., Campbell, M., Narechania, A., Majoros, W., Lewis, S., Thomas, P.D., and Ashburner, M. Assessment of genome-wide protein function classification for Drosophila melanogaster. *Genome Res.* 13:2118-28, 2003.

[28] Polakowska, R.R., Eickbush, T., Falciano, V., Razvi, F., and Goldsmith, L.A. Organization and evolution of the human epidermal keratinocyte transglutaminase I gene. *Proc Natl Acad Sci U S A*. 89:4476-80, 1992.

[29] Sahoo, S.S., Thomas, C., Sheth, A., York, W.S., and Tartir, S. Knowledge modeling and its application in life sciences: a tale of two ontologies. In *Proceedings of the 15th International Conference on World Wide Web* (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 317-326.

[30] Shimeld, S.M. Characterization of AmphiF-spondin reveals the modular evolution of chordate F-spondin genes. *Mol Biol Evol.* 15:1218-23, 1998.

[31] Sidhu, A.S., Dillon, T.S., et al., "Protein Ontology: Data Integration using Protein Ontology" in Database Modeling in Biology: Practices and Challenges. Z. Ma and J. Y. Chen. New York, NY, Springer Inc.:39 - 60, April 2006, ISBN: 0-387-30238-7.

[32] Smith, B., Ceuster, W., Klagger, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., and Rosse, C. Relations in Biomedical Ontologies *Genome Biol*. 6:R46, 2005.

[33] Weinstein, M., Yang, X., and Deng, C. Functions of mammalian Smad genes as revealed by targeted gene disruption in mice. *Cytokine Growth Factor Rev.* 11:49-58, 2000.

[34] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.-J., Mazumder, R., O'Donovan, C., Redaschi, N., and Suzek, B. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* 34:D187-191, 2006.

[35] Wu, C.H., Huang, H., Nikolskaya, A., Hu, Z., and Barker, W.C. The iProClass integrated database for protein functional analysis. *Comput Biol Chem* 28:87-96, 2004.

[36] Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.-S., Natale, D., Vinayaka, C.R., Hu, Z., Mazumder, R., Kumar, S., Kourtesi,s P., Ledley, R.S., Suzek, B.E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J.L., Chung, S., Castro-Alvear, J., Dinkov, G., and Barker, W.C. PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Res*. 32, D112-114, 2004.

[37] Yagi, K., Goto, D., Hamamoto, T., Takenoshita, S., Kato, M., and Miyazono, K. Alternatively spliced variant of Smad2 lacking exon 3. Comparison with wild-type Smad2 and Smad3. *J Biol Chem.* 274:703-9, 1999.

[38] Yamamoto, S., Asanuma, T., Takagi, T., and Fukuda, K.I. The molecule role ontology: an ontology for annotation of signal transduction pathway molecules in the scientific literature. *Comparative and Functional Genomics* 5:528 – 536, 2005.