PIR A UniProt CONSORTIUM MEMBER
Protein Information Resource

Introduction to Protein Ontology (PRO)

UNIVERSITY OF DELAWARE

PIR Protein Information Resource A UniProt CONSORTIUM MEMBER

ALZFORUM / PROTEIN ONTOLOGY KICK-OFF MEETING

OCTOBER 4-5, 2011 · BUFFALO, NY

http://www.bioontology.org/wiki/index.php/Alzforum_/_Protein_Ontology_Kick-Off_Meeting

Cathy H. Wu, Ph.D.

Director, Protein Information Resource (PIR)

Edward G. Jefferson Chair and Director

Center for Bioinformatics & Computational Biology, University of Delaware

Professor of Biochemistry & Molecular Biology, Georgetown University

PRO in OBO Foundry

Ontology for semantic integration of heterogeneous biological data

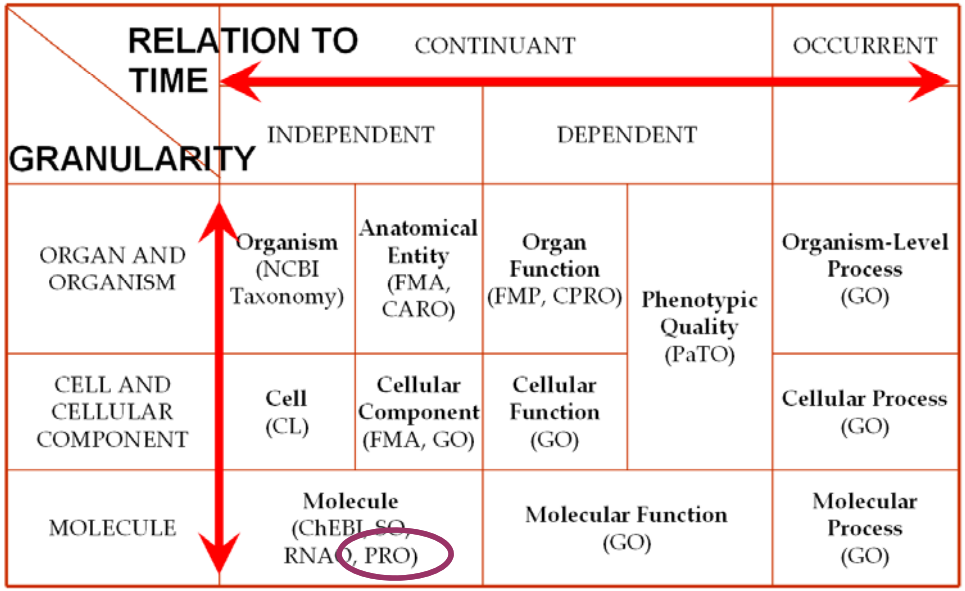
OBO Foundry

The Open Biological and Biomedical Ontologies

- Establishing a set of principles to create a suite of orthogonal interoperable reference ontologies
- First set of OBO Foundry ontologies

OBO Foundry ontologies

Title	Domain	Prefix	File
Biological process	biological process	GO	gene_ontology_edit.obo
Cellular component	anatomy	GO	gene_ontology_edit.obo
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo
Molecular function	biological function	GO	gene_ontology_edit.obo
Phenotypic quality	phenotype	PATO	quality.obo
PRotein Ontology (PRO)	proteins	PR	pro.obo
Xenopus anatomy and development	anatomy	XAO	xenopus_anatomy.obo
Zebrafish anatomy and development	anatomy	ZFA	zebrafish_anatomy.obo



Published online 8 October 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* **D539–D545**
doi:10.1093/nar/gkq907

The Protein Ontology: a structured representation of protein forms and complexes

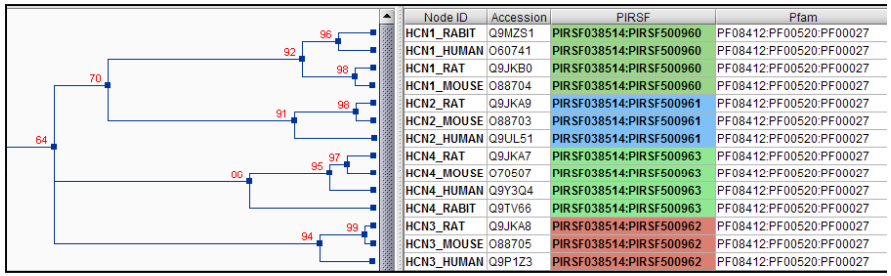
Darren A. Natale[1,*], Cecilia N. Arighi[2], Winona C. Barker[1], Judith A. Blake[3], Carol J. Bult[3], Michael Caudy[4], Harold J. Drabkin[3], Peter D'Eustachio[5], Alexei V. Evsikov[3], Hongzhan Huang[2], Jules Nchoutmboube[2], Natalia V. Roberts[2], Barry Smith[6], Jian Zhang[1] and Cathy H. Wu[1,2,*]

PRO Overview

PRO in OBO Foundry to represent protein entities

Three sub-ontologies to connect protein types necessary to model biology

- Ontology for Protein Evolution (ProEvo): Captures protein classes reflecting evolutionary relatedness of whole proteins



Node ID	Accession	PIRSF	Pfam
HCN1_RABIT	Q9MZS1	PIRSF038514:PIRSF500960	PF08412:PF00520:PF00027
HCN1_HUMAN	O60741	PIRSF038514:PIRSF500960	PF08412:PF00520:PF00027
HCN1_RAT	Q9JKB0	PIRSF038514:PIRSF500960	PF08412:PF00520:PF00027
HCN1_MOUSE	O88704	PIRSF038514:PIRSF500960	PF08412:PF00520:PF00027
HCN2_RAT	Q9JKA9	PIRSF038514:PIRSF500961	PF08412:PF00520:PF00027
HCN2_MOUSE	O88703	PIRSF038514:PIRSF500961	PF08412:PF00520:PF00027
HCN2_HUMAN	Q9UL51	PIRSF038514:PIRSF500961	PF08412:PF00520:PF00027
HCN4_RAT	Q9JKA7	PIRSF038514:PIRSF500963	PF08412:PF00520:PF00027
HCN4_MOUSE	Q70507	PIRSF038514:PIRSF500963	PF08412:PF00520:PF00027
HCN4_HUMAN	Q9Y3Q4	PIRSF038514:PIRSF500963	PF08412:PF00520:PF00027
HCN4_RABIT	Q9TV66	PIRSF038514:PIRSF500963	PF08412:PF00520:PF00027
HCN3_RAT	Q9JKA8	PIRSF038514:PIRSF500962	PF08412:PF00520:PF00027
HCN3_MOUSE	O88705	PIRSF038514:PIRSF500962	PF08412:PF00520:PF00027
HCN3_HUMAN	Q9P1Z3	PIRSF038514:PIRSF500962	PF08412:PF00520:PF00027

PR:000000676 *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel protein*	*family*
PR:000000705 *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 1*	*gene*
PR:000000706 *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2*	*gene*
PR:000000707 *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 3*	*gene*
PR:000000708 *potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4*	*gene*

- Ontology for Protein Forms (ProForm): Captures different protein forms of a given gene locus from genetic variations, alternative splicing, proteolytic cleavage, PTMs

PR:000002184 *Bcl2 antagonist of cell death*	*gene*	
PR:000002280 *Bcl2 antagonist of cell death isoform 1*	*sequence*	
PR:000003084 *Bcl2 antagonist of cell death isoform 1 phosphorylated form*	*modification*	*O35147; Q61337; Q92934*
PR:000003085 *Bcl2 antagonist of cell death isoform 1 phosphorylated 1*	*modification*	*O35147-1; Q61337-1*
PR:000003086 *Bcl2 antagonist of cell death isoform 1 phosphorylated 2*	*modification*	
PR:000003087 *Bcl2 antagonist of cell death isoform 1 phosphorylated 3*	*modification*	*Q61337-1:pS112/pS136*
PR:000003233 *Bcl2 antagonist of cell death isoform 1 phosphorylated 4*	*modification*	*Q61337-1:pS112/pS136/pS155*
PR:000003238 *Bcl2 antagonist of cell death isoform 1 phosphorylated 5*	*modification*	*O35147-1:pS112*
PR:000003269 *Bcl2 antagonist of cell death isoform 1 phosphorylated 6*	*modification*	*Q61337-1:pT201*
PR:000025849 *Bcl2 antagonist of cell death isoform 1 phosphorylated 7*	*modification*	*Q61337-1:pS136*
PR:000025850 *Bcl2 antagonist of cell death isoform 1 phosphorylated 8*	*modification*	*Q61337-1:pS128/pS136*

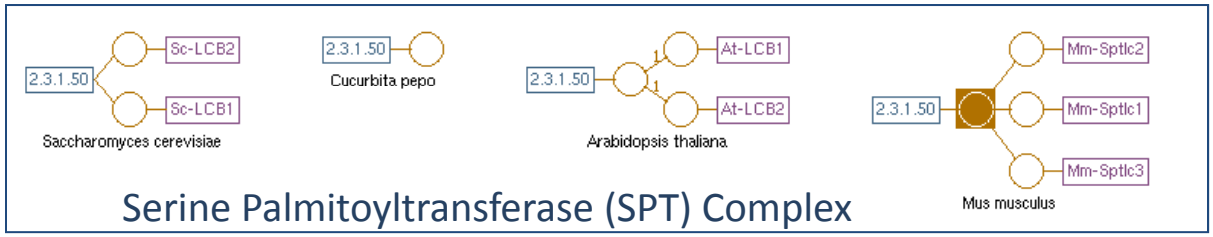
Need for Representing Proteins Forms

Alternative Splicing/Genetic Variation/PTM/Cleavage

Gene	Protein Form	Distinctive Attributes	
SMAD2	Long isoform phosphorylated (PR:000000468)	***NOT has_function*** GO:00036677 DNA binding	Function
	Short isoform phosphorylated (PR:000000469)	***has_function*** GO:00036677 DNA binding	
CUL1	Unmodified form (PR:000002507)	***NOT part_of*** GO:0019005 SCF ubiquitin ligase complex	Association
	Needylated form (PR:000000542)	***part_of*** GO:0019005 SCF ubiquitin ligase complex	
CD14	Membrane form (PR:000002149)	***located_in*** GO:0005886 plasma membrane	Localization
	Soluble form (PR:000002147)	***located_in*** GO:0005615 extracellular space	
ROCK1	Full length (PR:000002529)	***has_function*** GO:0004674 protein serine/threonine kinase activity	Modification
	Cleaved form (PR:000000563)	***Increased has_function*** GO:0004674 protein serine/threonine kinase activity	
CREBBP	Variant R -> P(1378) (PR:000000266)	**agent_in MIM:180849**, RUBINSTEIN-TAYBI SYNDROME SO:1000118, loss_of_function_of_polypeptide	Disease

PRO Overview

- Ontology for Protein Complexes (ProComp): Captures distinct complexes as they exist in different species and defines complexes through component proteins

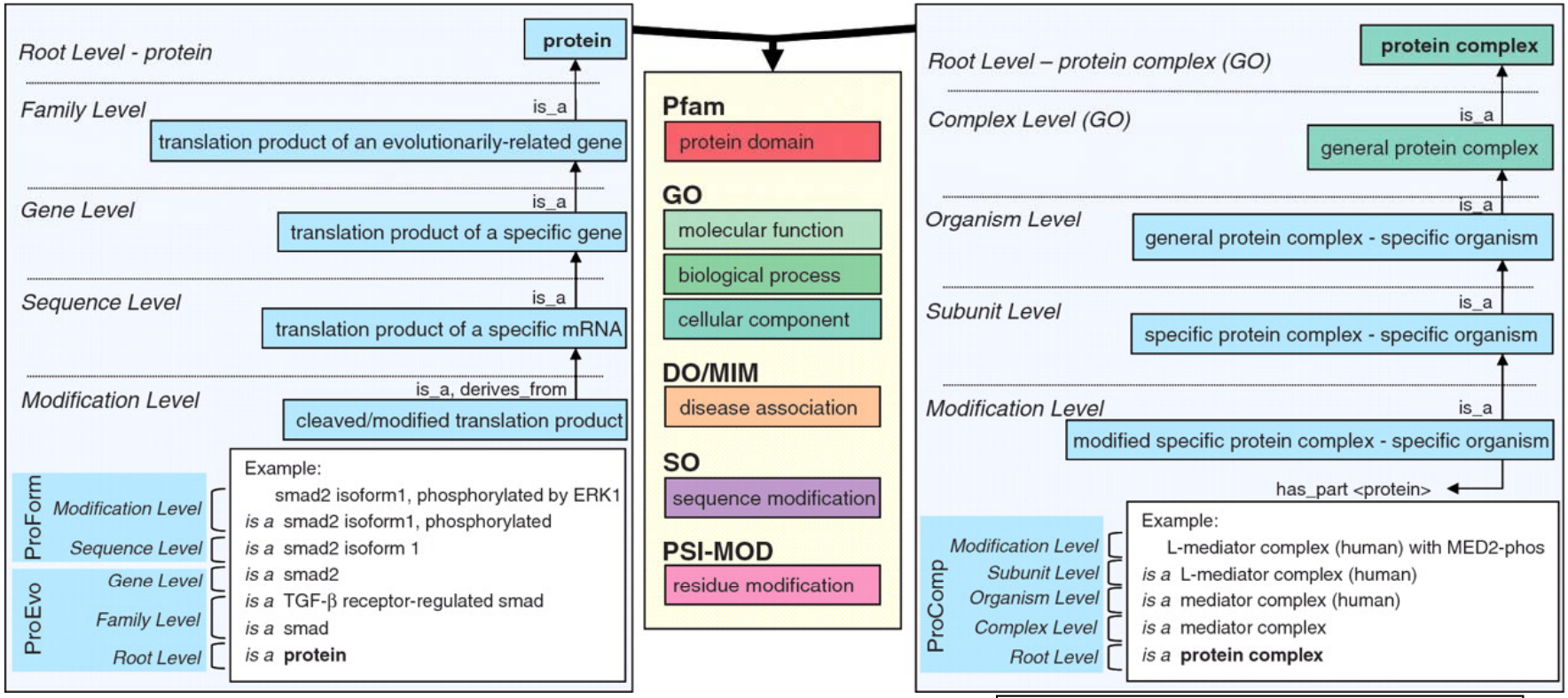


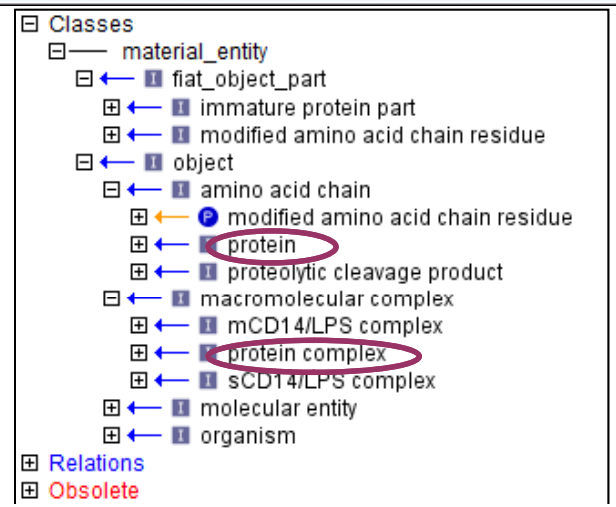
GO:0002178 *palmitoyltransferase complex*	
GO:0002179 *homodimeric serine palmitoyltransferase complex*	
PR:000026130 *bacterial serine palmitoyltransferase complex (Sphingomonas paucimobilis)*	*organism-complex*
PR:000026132 *bacterial serine palmitoyltransferase complex (Sphingomonas wittichii)*	*organism-complex*
PR:000026169 *bacterial serine palmitoyltransferase complex (Sphingobacterium multivorum)*	*organism-complex*
GO:0031211 *endoplasmic reticulum palmitoyltransferase complex*	
GO:0017059 *serine C-palmitoyltransferase complex*	
PR:000026146 *serine palmitoyltransferase complex 3 (human)*	*organism-complex*
PR:000026153 *serine palmitoyltransferase complex 5 (human)*	*organism-complex*
PR:000026145 *serine palmitoyltransferase complex core 2 (human)*	*organism-complex*

Why PRO?

- Provides formalization and precise annotation of specific protein classes/forms/ complexes, allowing accurate and consistent data mapping, integration and analysis
- Allows specification of relationships between PRO and other ontologies, such as GO, SO, PSI-MOD, MIM/Disease Ontology
- Provides stable unique identifiers to distinct protein types
- Provides a formal structure to support computer-based reasoning based on homology and shared protein attributes, including “ortho-isoform,” “ortho-modified form”

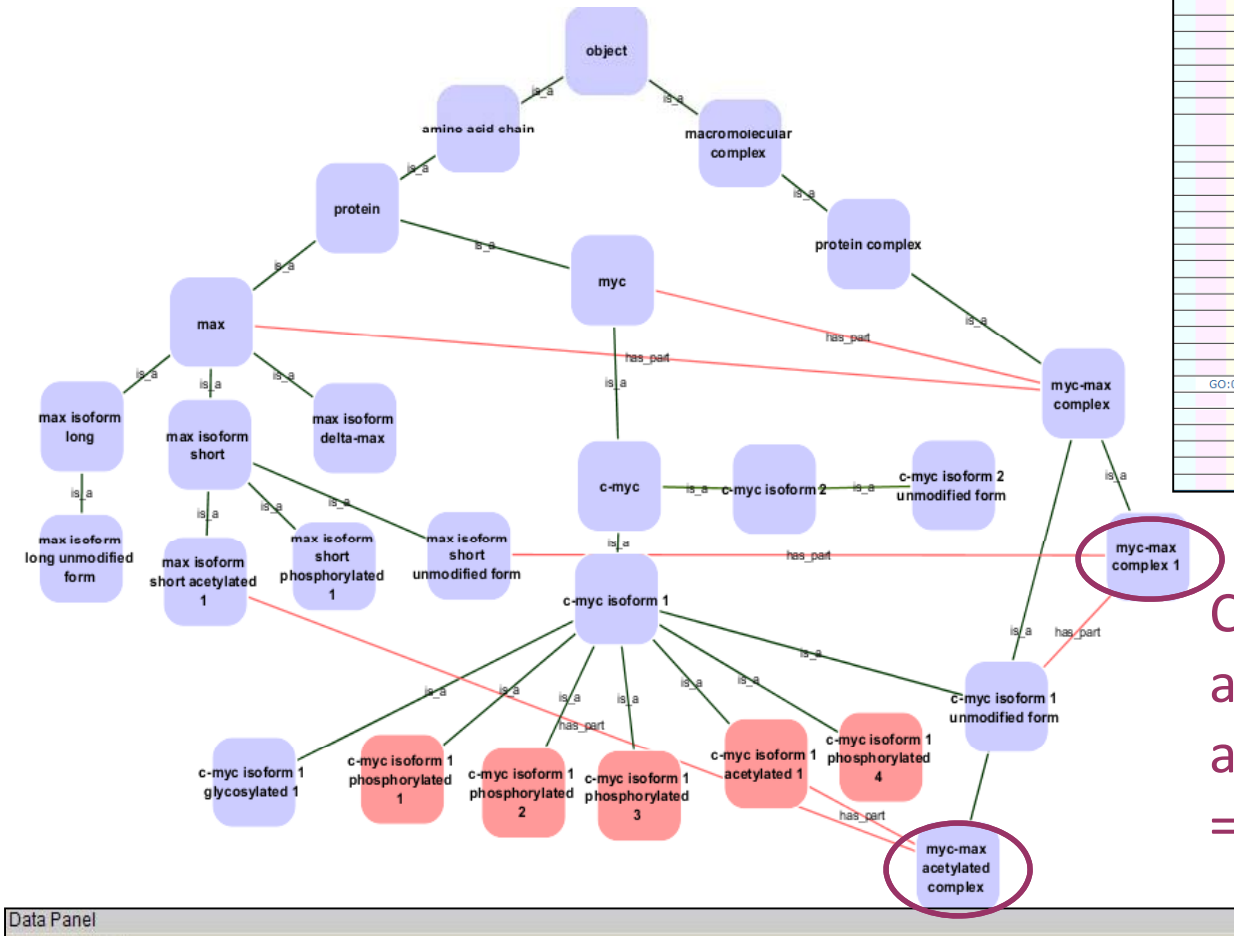
PRO Framework





- PRO (ProForm, ProEvo, ProComp) is aligned with other OBO Foundry ontologies under the umbrella of the Basic Formal Ontology (BFO)
- PRO terms are defined/annotated using other ontologies and resources via definition of relations or mappings when appropriate

PRO Network View



Object	Category
PR:000018263 ***amino acid chain***	
PR:000000001 ***protein***	
PR:000000020/PIRSF001705 *myc*	*family*
PR:000000084 *c-myc*	*gene*
UniProtKB:P01106 *c-myc (human)*	*organism-gene*
UniProtKB:P01108 *c-myc (mouse)*	*organism-gene*
PR:000000229 *c-myc isoform 2*	*sequence*
UniProtKB:P01106-2 *c-myc isoform 2 (human)*	*organism-sequence*
UniProtKB:B2RSN1 *c-myc isoform 2 (mouse)*	*organism-sequence*
PR:000000228 *c-myc isoform 1*	*sequence*
UniProtKB:P01106-1 *c-myc isoform 1 (human)*	*organism-sequence*
UniProtKB:P01108-1 *c-myc isoform 1 (mouse)*	*organism-sequence*
PR:000000535 *c-myc isoform 1 acetylated 1*	*modification*
PR:000026033 *c-myc isoform 1 acetylated 1 (human)*	*organism-modification*
PR:000000536 *c-myc isoform 1 glycosylated 1*	*modification*
PR:000026051 *c-myc isoform 1 glycosylated 1 (human)*	*organism-modification*
PR:000000537 *c-myc isoform 1 phosphorylated 1*	*modification*
PR:000026052 *c-myc isoform 1 phosphorylated 1 (mouse)*	*organism-modification*
PR:000000538 *c-myc isoform 1 phosphorylated 2*	*modification*
PR:000026053 *c-myc isoform 1 phosphorylated 2 (mouse)*	*organism-modification*
PR:000026054 *c-myc isoform 1 phosphorylated 2 (human)*	*organism-modification*
PR:000000539 *c-myc isoform 1 phosphorylated 3*	*modification*
PR:000026055 *c-myc isoform 1 phosphorylated 3 (human)*	*organism-modification*
PR:000026040 *c-myc isoform 1 phosphorylated 4*	*modification*
PR:000026041 *c-myc isoform 1 phosphorylated 4 (human)*	*organism-modification*
PR:000002503 *c-myc isoform 1 unmodified form*	*modification*
PR:000026038 *c-myc isoform 1 unmodified form (human)*	*organism-modification*
PR:000026056 *c-myc isoform 1 unmodified form (mouse)*	*organism-modification*
GO:0032991 ***macromolecular complex***	
GO:0043234 ***protein complex***	
GO:0071943 *myc-max complex*	*complex*
PR:000026036 *myc-max complex 1*	*complex*
PR:000026047 *myc-max complex 1 (human)*	*organism-complex*
PR:000026034 *myc-max complex acetylated*	*complex*
PR:000026035 *myc-max complex acetylated (human)*	*organism-complex*

Connecting protein forms and complexes with annotation
=> Model biology/disease

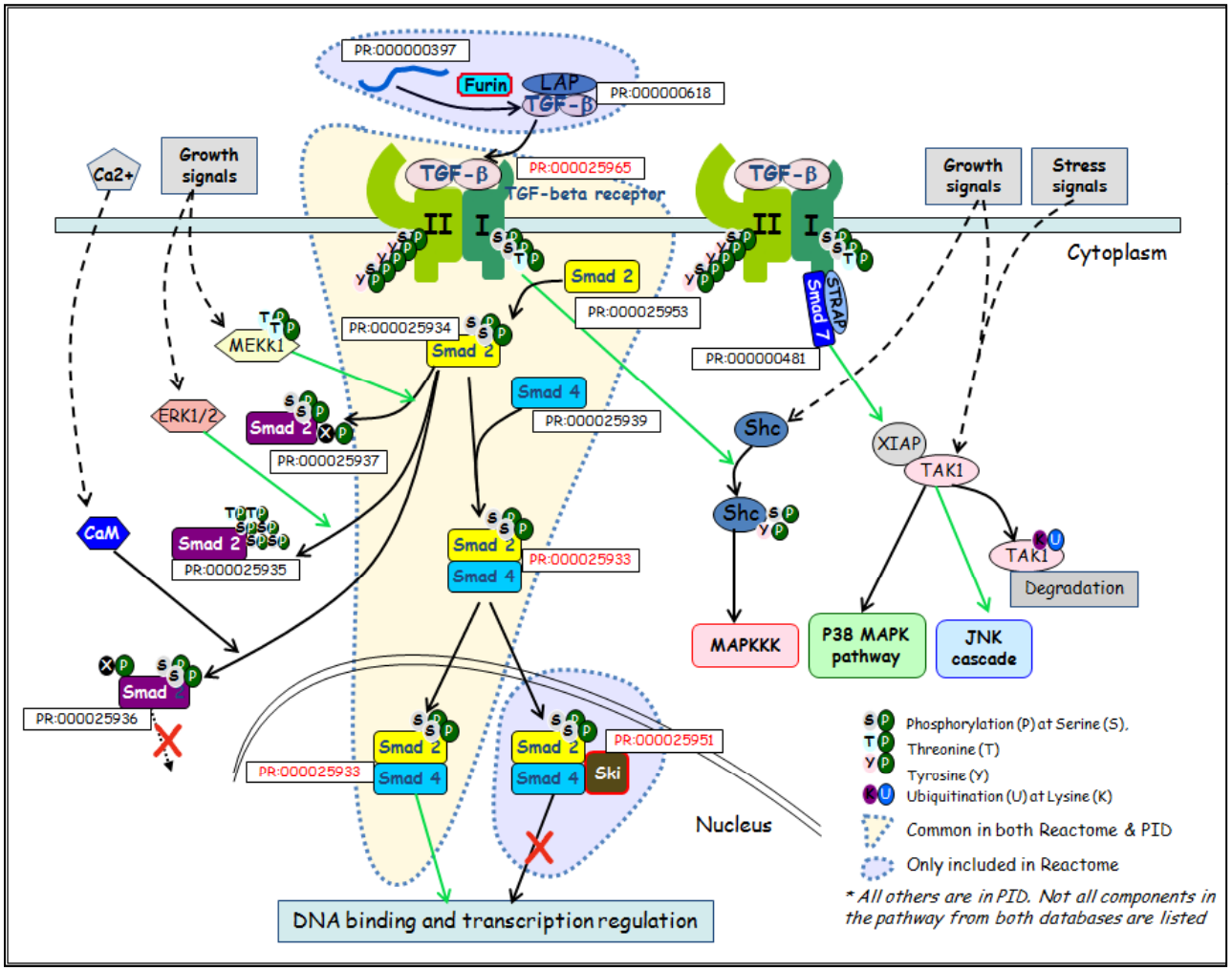
Data Panel

ID	Column 2	Column 3	Column 4	Column 5	Column 6
PR:000000535	c-myc protein isoform 1 acetylated 1	participates_in	GO:0045941	positive regulation of transcription	UniProtKB:P01106-1,acK143/acK157/acK275/acK317/acK323/acK371
PR:000000538	c-myc protein isoform 1 phosphorylated 2	participates_in	GO:0016481	negative regulation of transcription	UniProtKB:P01108-1, pT58/pS62, UniProtKB:P01106-1, pT58/pS62
PR:000000539	c-myc protein isoform 1 phosphorylated 3	participates_in	GO:0045766	positive regulation of angiogenesis	UniProtKB:P01106-1, pT58
PR:000026040	c-myc protein isoform 1 phosphorylated 4	participates_in	GO:0016481	negative regulation of transcription	UniProtKB:P01106-1, pS62/pS71
PR:000000537	c-myc protein isoform 1 phosphorylated 1	participates_in	GO:0045941	positive regulation of transcription	UniProtKB:P01108-1, pS62

PRO in Pathway Context

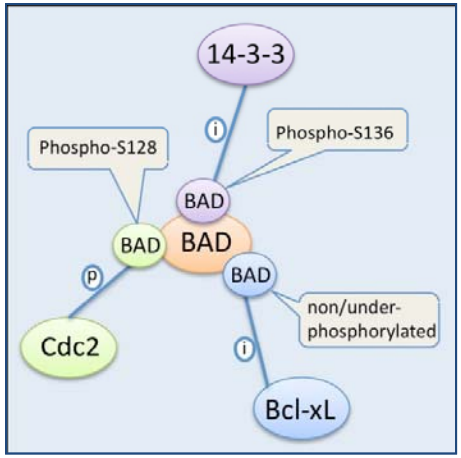
- Representation of protein forms & complexes in biological/pathway/network context
- Connecting multiple pathway/network resources

TGF-β Signaling Pathway: *Reactome, PID, coupled with literature curation*

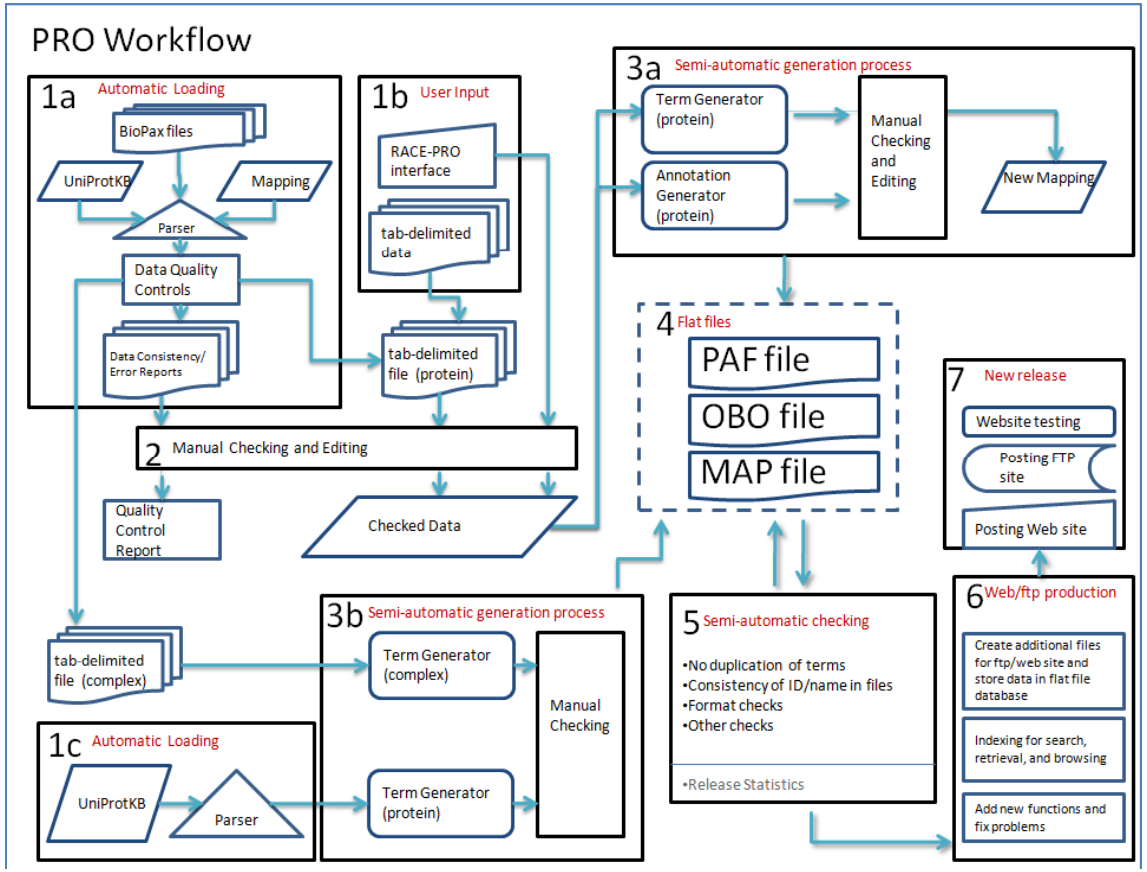


PRO Workflow

- Data Sources
  - Manual annotation (curator, user): sourceforge tracker and Race-PRO
  - Semi-automated processing of external databases (e.g., UniProtKB, Reactome, MouseCyc, EcoCyc)
- Integration with text mining tool: eFIP (*Functional Impact of Phosphorylation)*
- Distribution Files
  - Ontology (OBO)
  - Annotation (PAF)
  - Mappings (exact; is_a)



eFIP text mining



PRO Dissemination

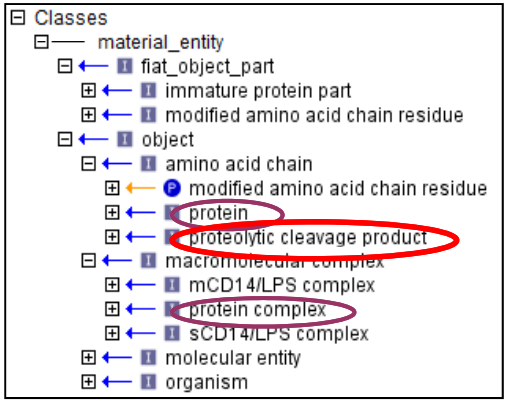
- PRO Website (http://pir.georgetown.edu/pro)
  - Searching, browsing, downloading
- PRO Views
  - Entry view
  - Table summary
  - OBO stanza, OWL
  - Ontology hierarchy
  - Cytoscape network
- PRO Link: Persistent URL: http://purl.obolibrary.org/obo/PR_xxxxxxxxx
- OBO Foundry (http://www.obofoundry.org/)
- NCBO Bioportal (http://bioportal.bioontology.org/)

PRO Communities

- Ontology Developers
  - GO ontology: Interfaces of GO/PRO complexes; GO definition (e.g., GO:0005109)
  - GO annotation: precise annotation of protein forms in PomBase
  - Dendritic Cell Ontology: Define cell types based on +/- protein types [PMID:19243617]
  - Annotation Ontology for annotating scientific documents on the web [PMID:21624159]
- Semantic Resources
  - Royal Society of Chemistry (RSC); Science Collaboration Framework; Semantic Web Applications in Neuromedicine (SWAN); Neuroscience Information Framework (NIF)
- Pathway/Process-Modeling Resources:
  - Reactome, MouseCyc, EcoCyc/BioCyc, Center for Molecular Immunology (Duke)
- Molecule-Modeling Resources: Int'l Union of Basic and Clinical Pharmacology (IUPhar)
- Pharma/Clinical Communities: Drug Discovery & Disease Biomarker
  - Alzforum
  - Salivaomics KB/SALO (Saliva Ontology): Saliva Biomarkers
  - Pistoia

AlzForum Driving Project

- Aids AlzForum as a comprehensive source of information about Alzheimer's disease (AD), linking to the wider biomedical knowledge via community-accepted ontologies
- Clinical driving project to address the needs of AD researchers: identify gaps in PRO for representing AD-related protein entities, especially those involved in etiology
- AD-related protein entities include various types of proteolytic cleavage products and protein complexes: Drives new terms/relationships: (i) proteolytic cleavage product (PR:000018264); (ii) *union_of* terms (PR:000025744 = set of 2 precursors)
- Current Coverage: (i) amyloid beta A4 protein (APP) (variants and cleavage products): 86 terms; (ii) microtubule-associated protein Tau (mutations and population variants): 23 terms; (iii) secretases (complexes): 10 terms



expand	sort (ID)	sort (STR)	find	Category
PR:000018263 ***amino acid chain***				
PR:000000001 ***protein***				
PR:000004168 *amyloid beta A4 protein*				*gene*
PR:000019036[d] *amyloid beta A4 protein proteolytic cleavage product*				*modification*
PR:000025744 *precursor of amyloid beta A4 protein gamma-secretase C-terminal fragment*				*union*
PR:000025589[d] *gamma-secretase C-terminal fragment 50*				*modification*
PR:000025585[d] *gamma-secretase C-terminal fragment 57*				*modification*
PR:000025586[d] *gamma-secretase C-terminal fragment 59*				*modification*

PRO Consortium Team *(current)*

Protein Information Resource (PIR) [Georgetown U & U Delaware]

Cathy Wu, Cecilia Arighi, Darren Natale, Natalia Roberts
Hongzhan Huang, Jian Zhang

The Jackson Lab – Mouse Genome Informatics (MGI)

Judith Blake, Carol Bult, Harold Drabkin, Alexei Evsikov

University at Buffalo-SUNY

Barry Smith, Alan Ruttenberg, Alexander Diehl

NYU School of Medicine – Reactome

Peter D'Eustachio, Michael Caudy

AlzForum

Elizabeth Wu

National Institute of General Medical Sciences

1R01GM080646-01
3R01GM080646-04S2
2R01GM080646-06