The Protein Ontology: An Evolution

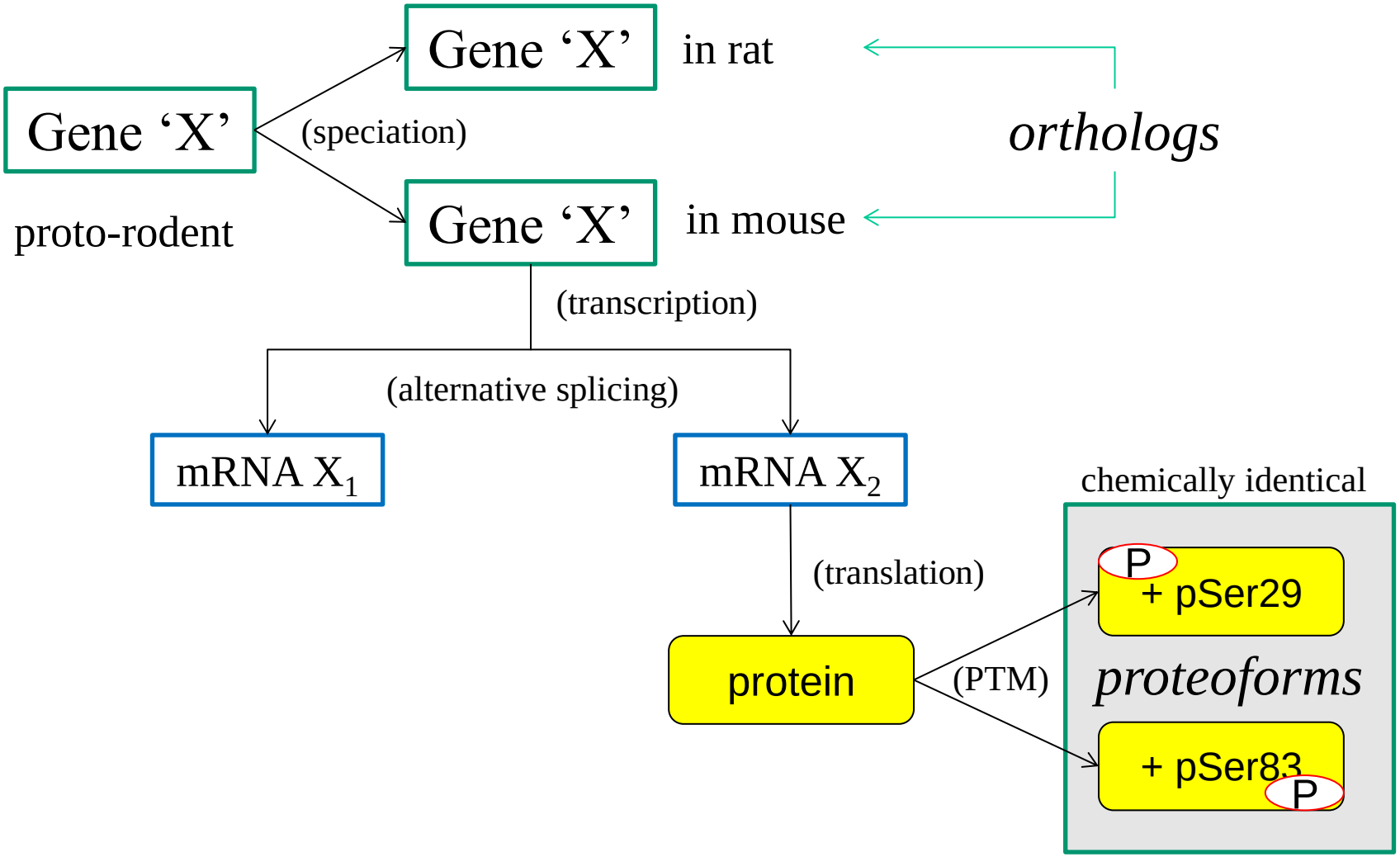
251st ACS National Meeting & Exposition

Chemistry, Data & the Semantic Web: An Important Triple to Advance Science

San Diego, CA

Darren A. Natale, Ph.D.
Protein Science Team Lead, PIR
Research Assistant Professor, GUMC

Some Background



Why create PRO:
Provide precise targets for annotation

Mothers against decapentaplegic homolog 2

Smad 2

GO annotation of SMAD2_HUMAN:

Cellular Component:

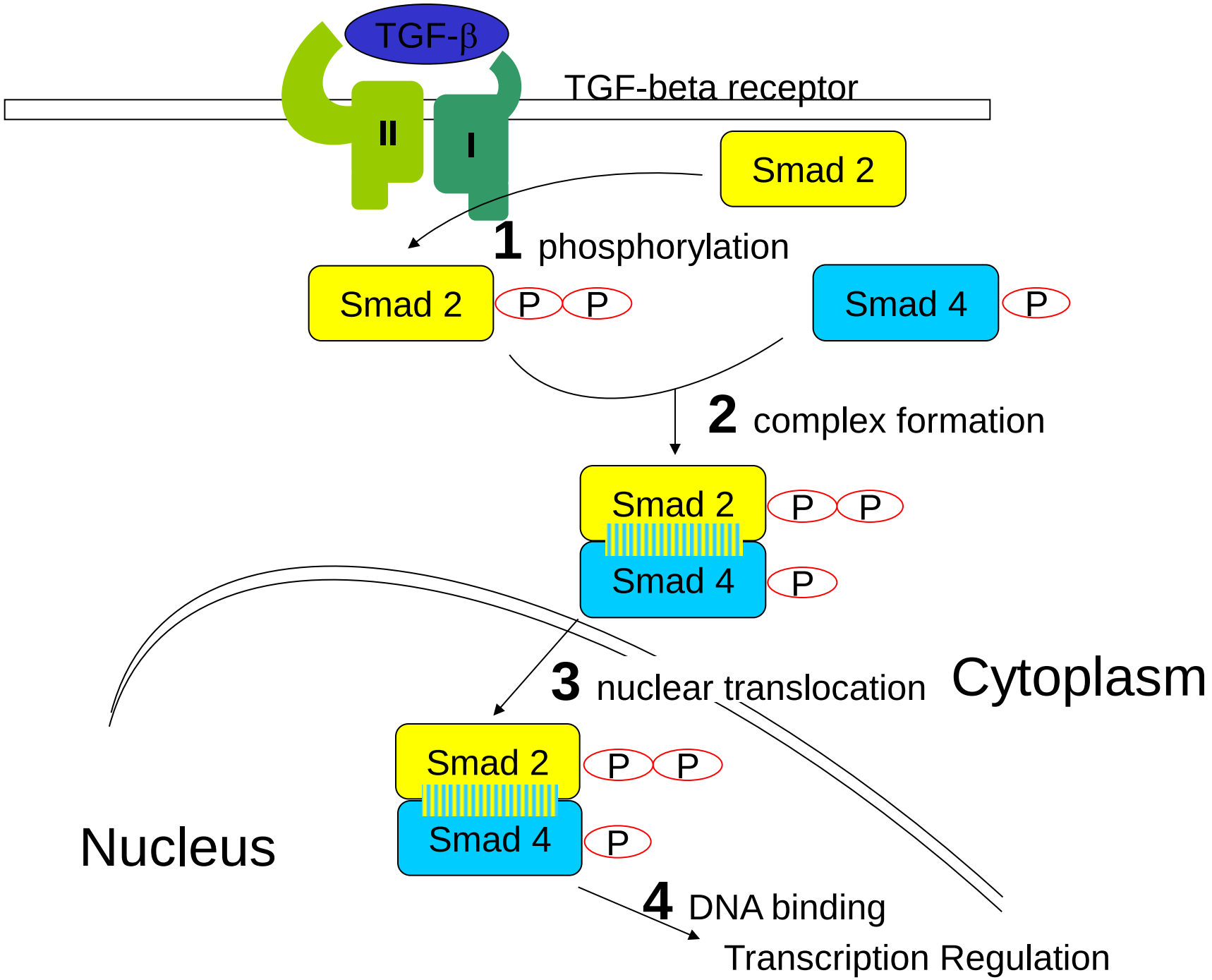
- nucleus

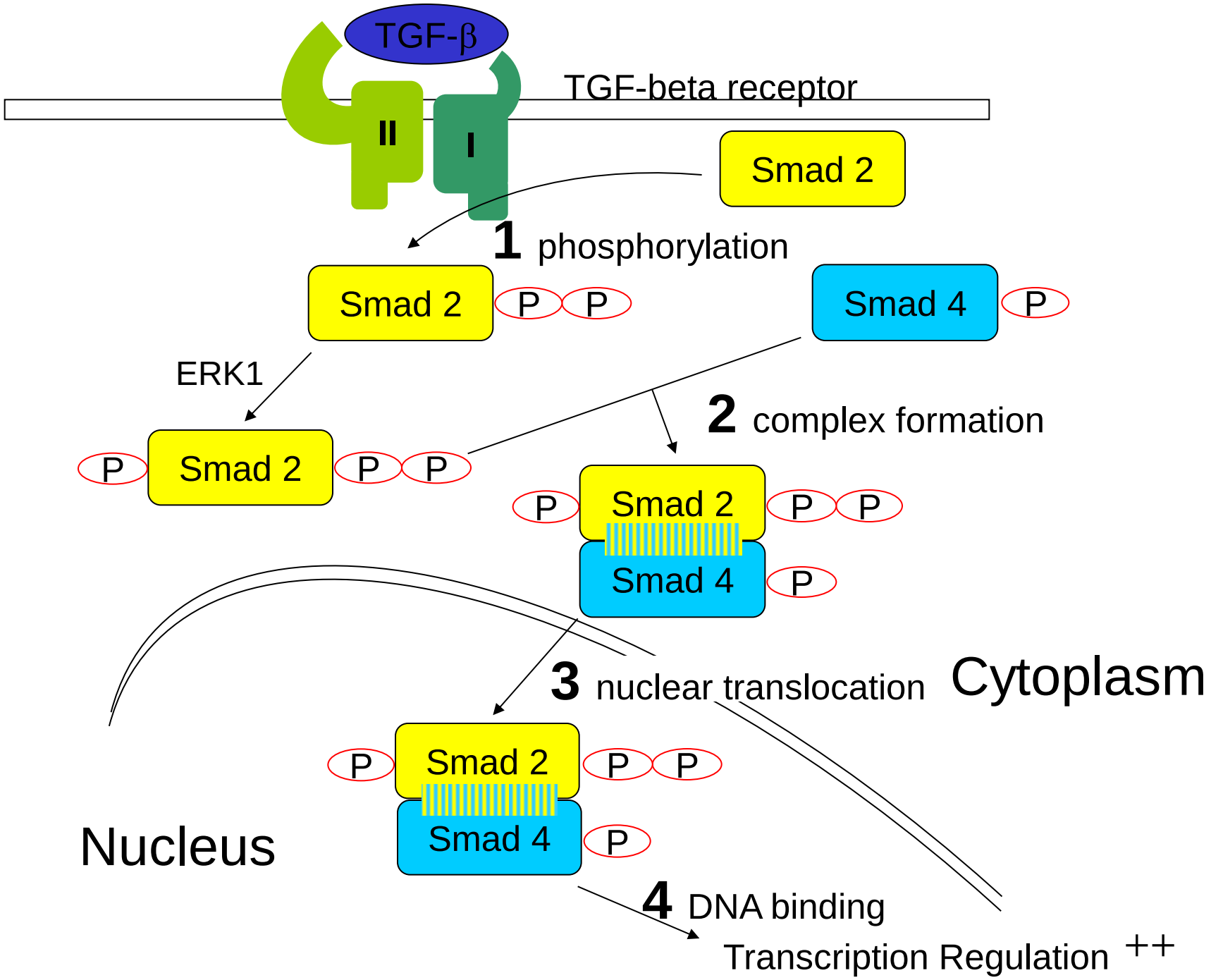
Molecular Function:

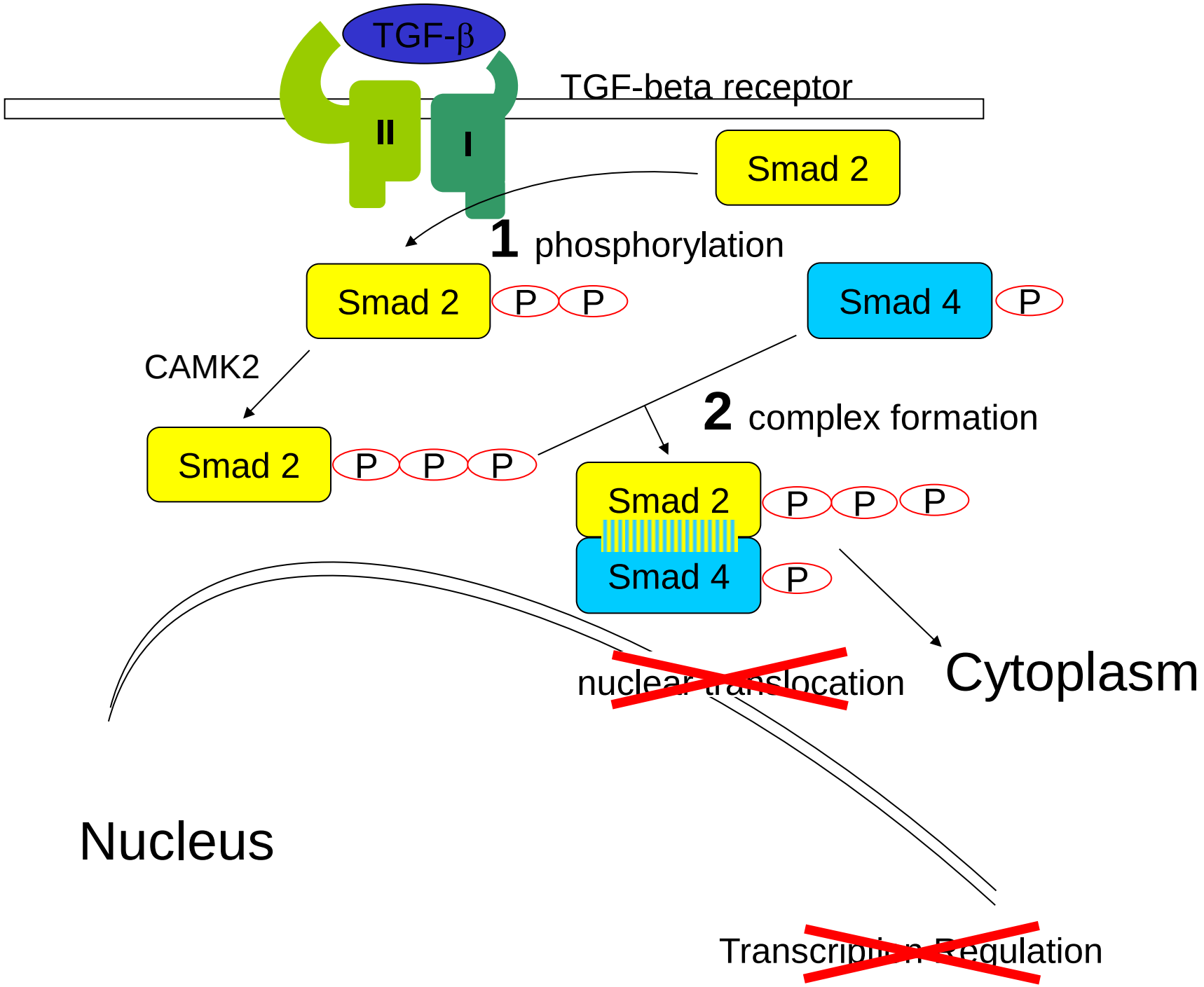
- protein binding

Biological Process:

- signal transduction
- regulation of transcription, DNA-dependent


TGF-β
TGF-beta receptor
II
I
Smad 2
1 phosphorylation
Smad 2
P
P
Smad 4
P
2 complex formation
Smad 2
P
P
Smad 4
P
3 nuclear translocation
Cytoplasm
Nucleus
Smad 2
P
P
Smad 4
P
4 DNA binding
Transcription Regulation


TGF-β
TGF-beta receptor
II
I
Smad 2
1 phosphorylation
Smad 2
P P
Smad 4
P
ERK1
2 complex formation
P
Smad 2
P P
P
Smad 2
P P
Smad 4
P
3 nuclear translocation
Cytoplasm
Nucleus
P
Smad 2
P P
Smad 4
P
4 DNA binding
Transcription Regulation ++


TGF-β
TGF-beta receptor
II
I
Smad 2
1 phosphorylation
Smad 2
P P
Smad 4
P
CAMK2
2 complex formation
Smad 2
P P P
Smad 2
P P P
Smad 4
P
nuclear translocation
Cytoplasm
Nucleus
Transcription Regulation

Smad 2	"normal"	•Cytoplasmic	SMAD2_HUMAN
Smad 2 P P	TGF-β receptor phosphorylated	•Forms complex •Nuclear •Txn upregulation	SMAD2_HUMAN
Smad 2 P P P	ERK1 phosphorylated	•Forms complex •Nuclear •Txn upregulation++	SMAD2_HUMAN
P Smad 2 P P	CAMK2 phosphorylated	•Forms complex •Cytoplasmic •No Txn upregulation	SMAD2_HUMAN
Smad 2	alternatively spliced short form	•Cytoplasmic	SMAD2_HUMAN
Smad 2 P P	phosphorylated short form	•Nuclear •Txn upregulation	SMAD2_HUMAN
Smad 2 X	point mutation (causative agent: large intestine carcinoma)	•Doesn't form complex •Cytoplasmic •No Txn upregulation	SMAD2_HUMAN

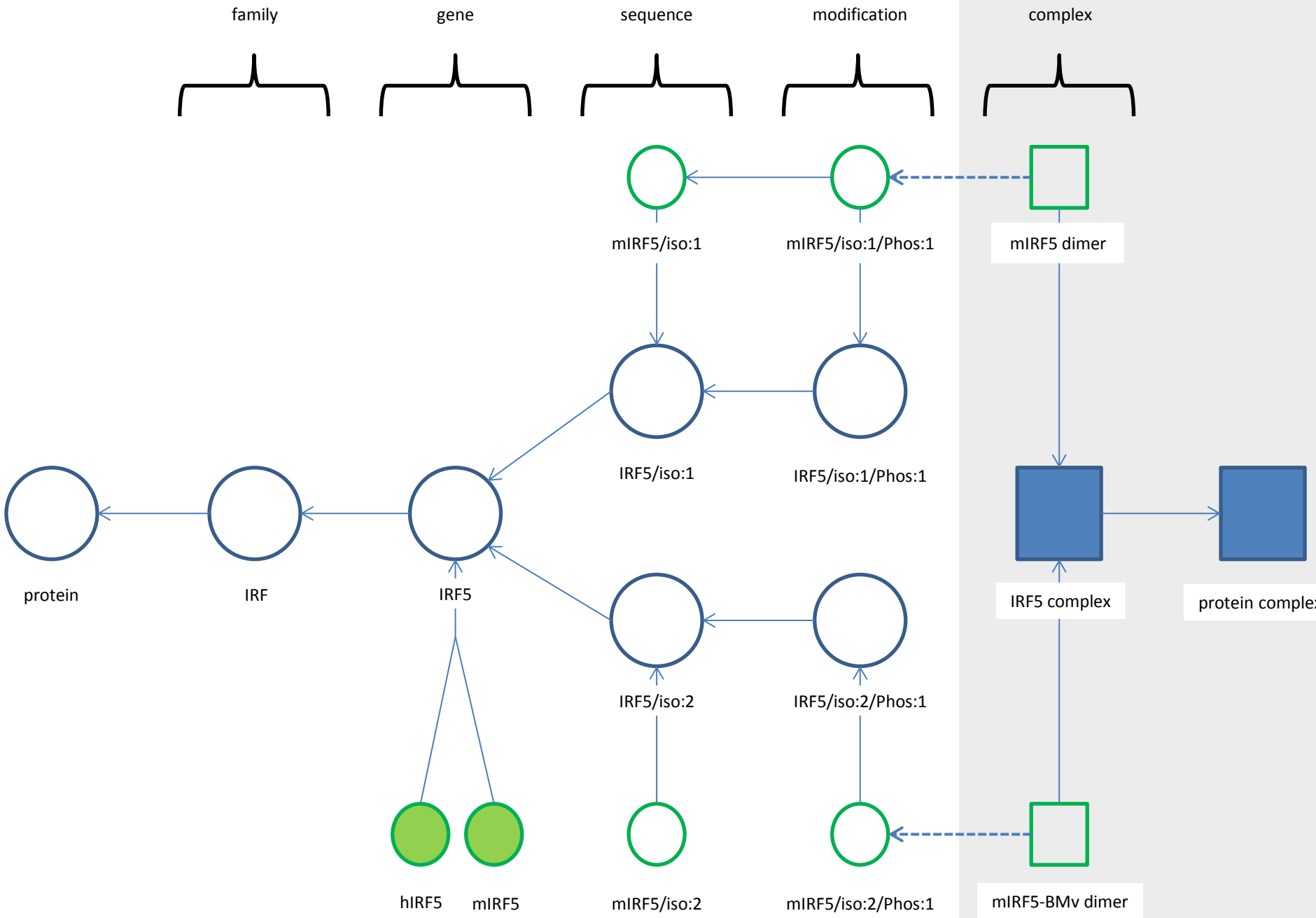
Smad 2	"normal"	•Cytoplasmic	PR:00000468
Smad 2 P P	TGF-β receptor phosphorylated	•Forms complex •Nuclear •Txn upregulation	PR:00000650
Smad 2 P P P	ERK1 phosphorylated	•Forms complex •Nuclear •Txn upregulation++	PR:00000651
P Smad 2 P P	CAMK2 phosphorylated	•Forms complex •Cytoplasmic •No Txn upregulation	PR:00000652
Smad 2	alternatively spliced short form	•Cytoplasmic	PR:00000469
Smad 2 P P	phosphorylated short form	•Nuclear •Txn upregulation	PR:00000656
Smad 2 X	point mutation (causative agent: large intestine carcinoma)	•Doesn't form complex •Cytoplasmic •No Txn upregulation	PR:00000470

Sometimes being precise means you need ***broader*** categories

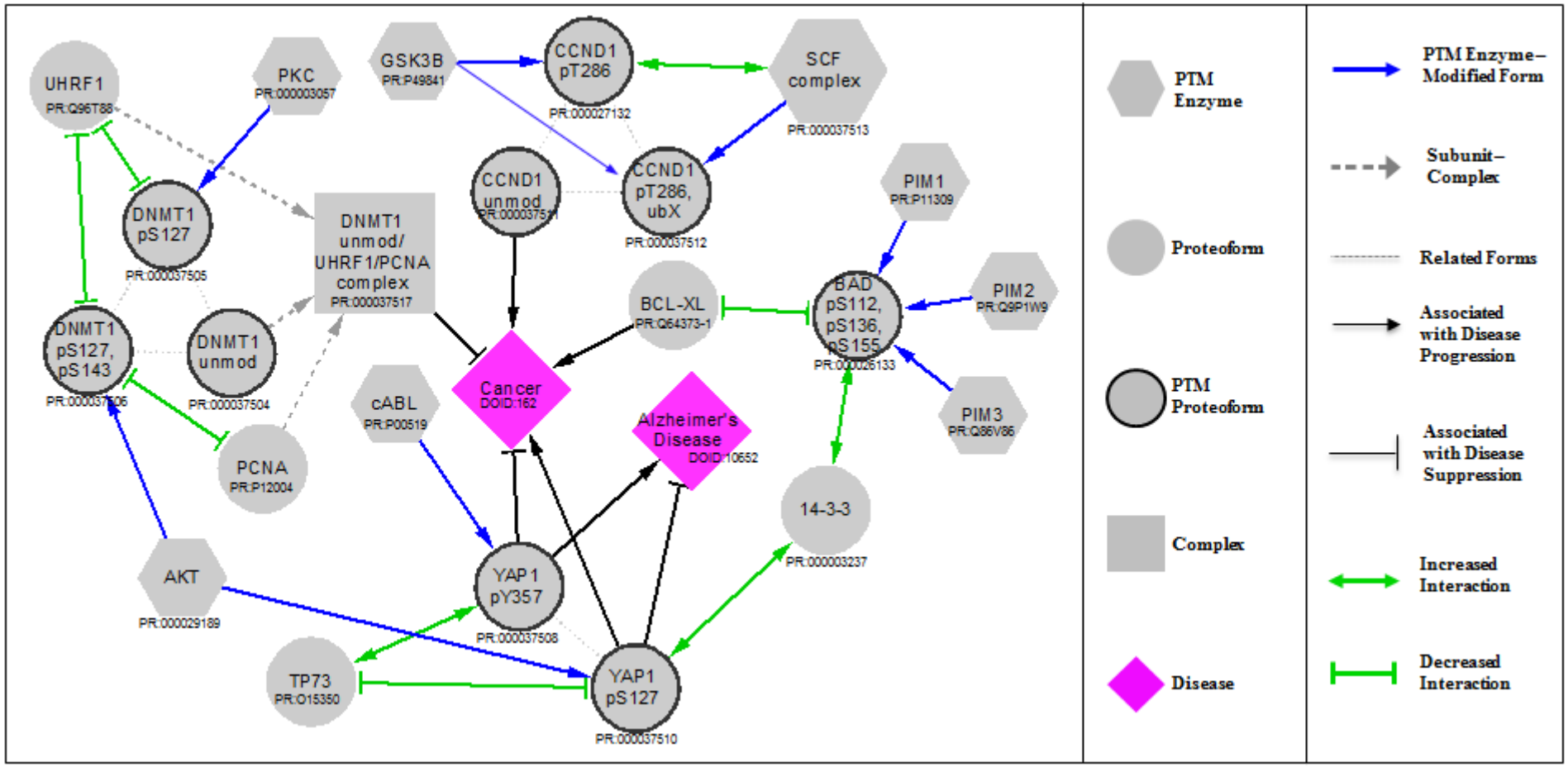
BioMarker	Alternative Names	Clone	PRO	Scope
CD45	PTPRC, B220, gp180, LCA, leukocyte common antigen, Ly-5, Lyt-4, lymphocyte antigen 5, T200	HI30	PR:000001006	All forms of CD45 (human+)
CD45RA		HI100	PR:000001015	isoform RA (human+)
CD45RB		MT4 (6B6)	PR:000001016	isoform RB (human+)
CD45RO		UCHL1	PR:000001017	isoform RO (human+)
Src	SRC1, ASV, c-src	3828	PR:P12931	All forms of Src (human)
Src (pY418)		K98-37	PR:000027237	Any Src phosphorylated on Y418 (human)

"Smad2 Is Essential for Maintenance of the Human and Mouse Primed Pluripotent Stem Cell State"

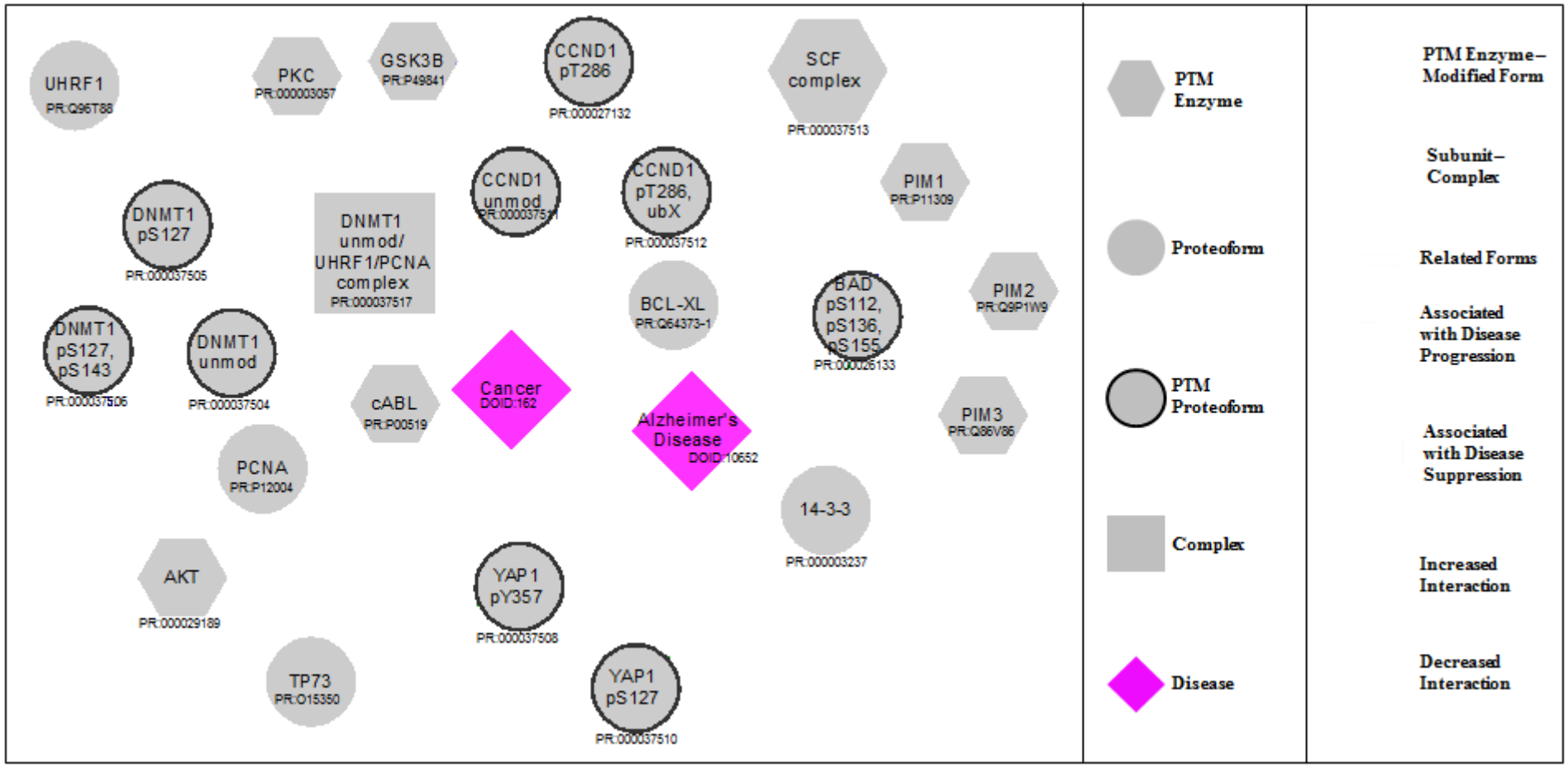
- some exp'ts done using hSmad2 (PR:Q15796)
- other exp'ts done using mSmad2 (PR:Q62432)
- conclusion meant to apply to both (PR:000000364)


family
gene
sequence
modification
complex
mIRF5/iso:1
mIRF5/iso:1/Phos:1
mIRF5 dimer
IRF5/iso:1
IRF5/iso:1/Phos:1
protein
IRF
IRF5
IRF5 complex
protein complex
IRF5/iso:2
IRF5/iso:2/Phos:1
hIRF5
mIRF5
mIRF5/iso:2
mIRF5/iso:2/Phos:1
mIRF5-BMv dimer

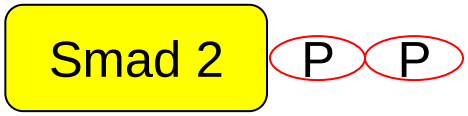
But Why An Ontology?



But Why An Ontology?



Ontology provides opportunity for rich relations

PRO, evolving:
How to represent  ?

Good:
= **PR:000000468** ! smad 2 isoform 1
+ *has_part* **MOD:00696** ! phosphorylated residue

Better:
= **PR:000000468** ! smad 2 isoform 1
+ *has_part* **MOD:00046** ! O-phospho-L-serine

Best:
= **PR:000000468** ! smad 2 isoform 1
+ *has_part* **MOD:00046** ! O-phospho-L-serine [at position 465]
+ *has_part* **MOD:00046** ! O-phospho-L-serine [at position 467]

Future: PRO “sites” sub-ontology for computable position information

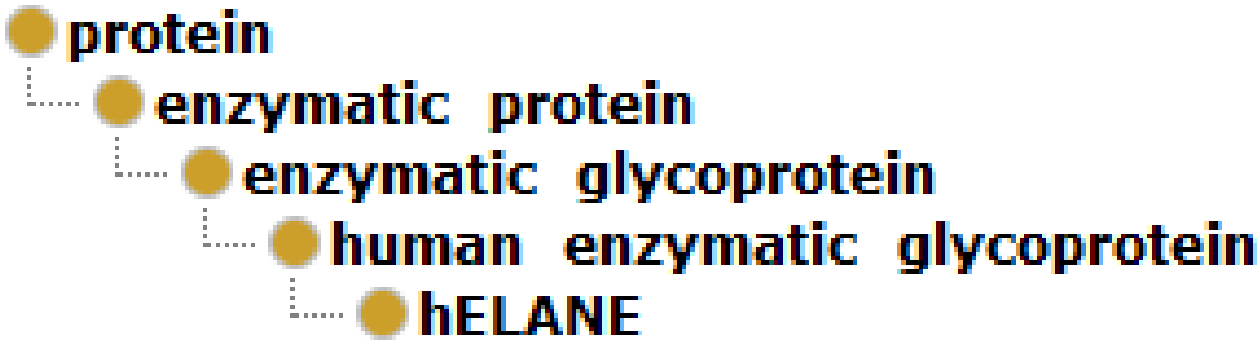
Toward A Better Way Of Building

protein

enzymatic protein = protein w/catalytic activity

enzymatic glycoprotein = enzymatic protein w/glycan

human enzymatic glycoprotein = enzymatic glycoprotein in human



human neutrophil elastase (*h*ELANE): a protein in human that has peptidase activity and is post-translationally modified with N-acetylglucosamine (GlcNAc)

Reformulate…

human enzymatic glycoprotein

=	*is_a*	protein
+	*capable_of*	catalytic activity
+	*has_part*	glycan
+	*in_taxon*	*Homo sapiens*

*h*ELANE	*is_a*	protein
*h*ELANE	*capable_of*	peptidase activity
*h*ELANE	*has_part*	GlcNAc
*h*ELANE	*in_taxon*	*Homo sapiens*

Reasoning…

Examine statements to (automatically) find new relationships or detect conflicts

The Benefit, Demonstrated

human enzymatic glycoprotein

=	*is_a*	protein
+	*capable_of*	catalytic activity
+	*has_part*	glycan
+	*in_taxon*	*Homo sapiens*

*h*ELANE *is_a*	protein		✓
*h*ELANE *capable_of*	peptidase activity	(→ *is_a* catalytic...)	✓
~~*h*ELANE *has_part*~~	~~GlcNAc~~	~~(→ *is_a* glycan)~~	
*h*ELANE *in_taxon*	*Homo sapiens*		✓

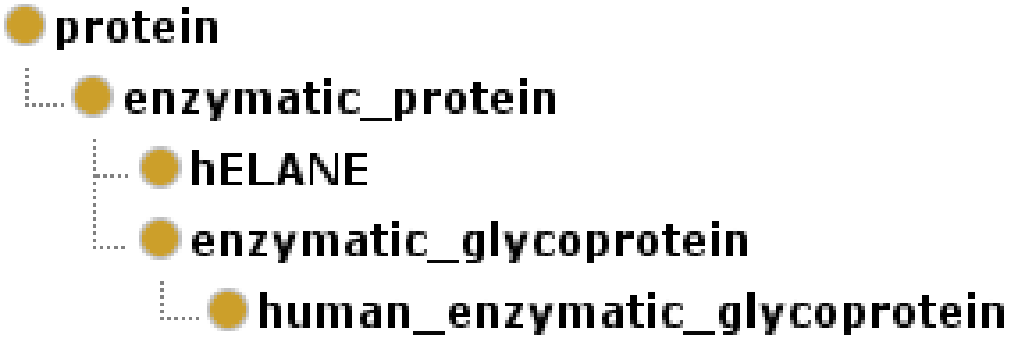
What if it's discovered that *h*ELANE isn't glycosylated?

The Benefit, Demonstrated

human enzymatic glycoprotein

=	*is_a*	protein
+	*capable_of*	catalytic activity
+	*has_part*	glycan
+	*in_taxon*	*Homo sapiens*

*h*ELANE *is_a*	protein		✓
*h*ELANE *capable_of*	peptidase activity	(→ *is_a* catalytic...)	✓
~~*h*ELANE *has_part*~~	~~GlcNAc~~	~~(→ *is_a* glycan)~~	
*h*ELANE *in_taxon*	*Homo sapiens*		✓



The Benefit, Demonstrated

~~human enzymatic glycoprotein~~

= *is_a* protein

~~catalytic activity~~

~~glycan~~

\+ *in_taxon* *Homo sapiens*

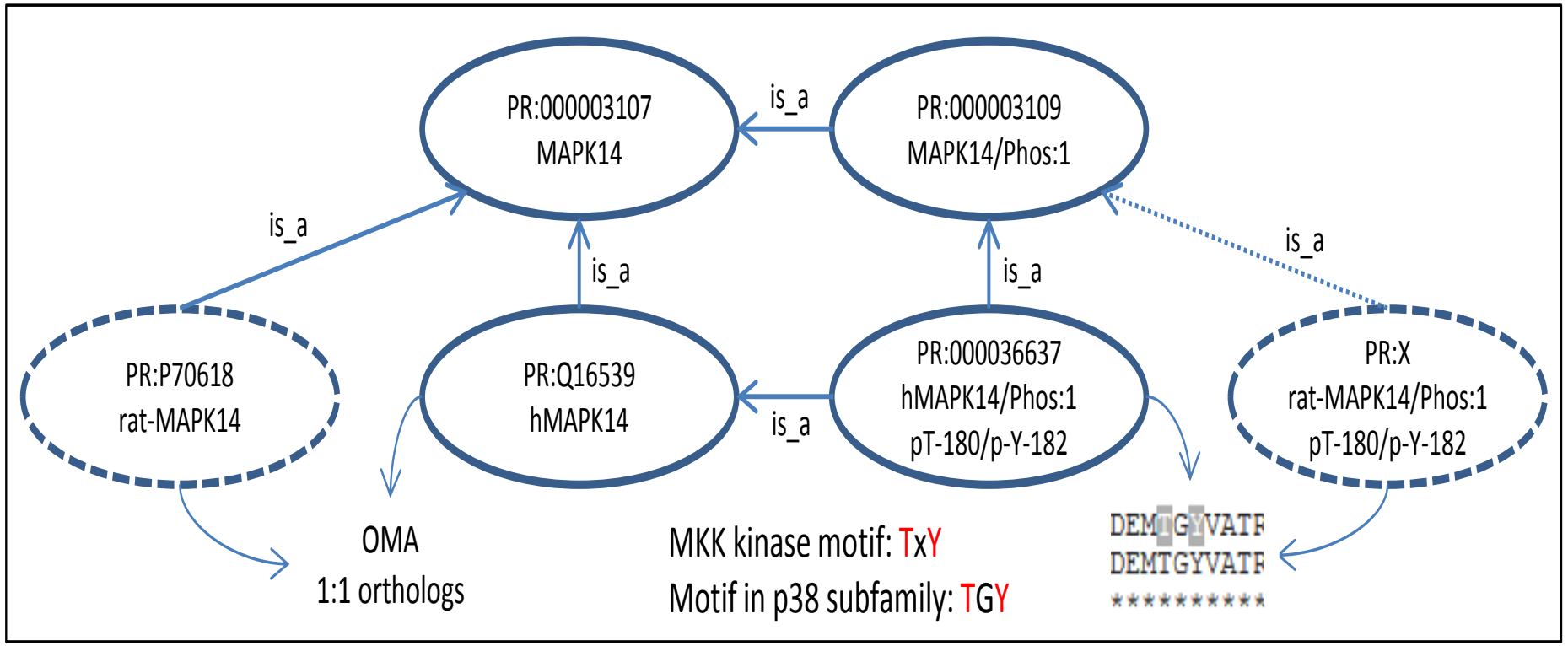
What if you've made a statement elsewhere that glycoproteins are never enzymes?

A reasoner would flag the term as inconsistent with rules

Lessons Learned…

- It is better to make statements on a more elemental level than make statements that are more complex
  - can add and remove as needed
  - easier to understand
  - enables reasoning
- It is better to make statements about evidence rather than make statements about conclusions
  - can import needed information
- It is beneficial to note mutually-exclusive conditions
  - helps flag inconsistent cases

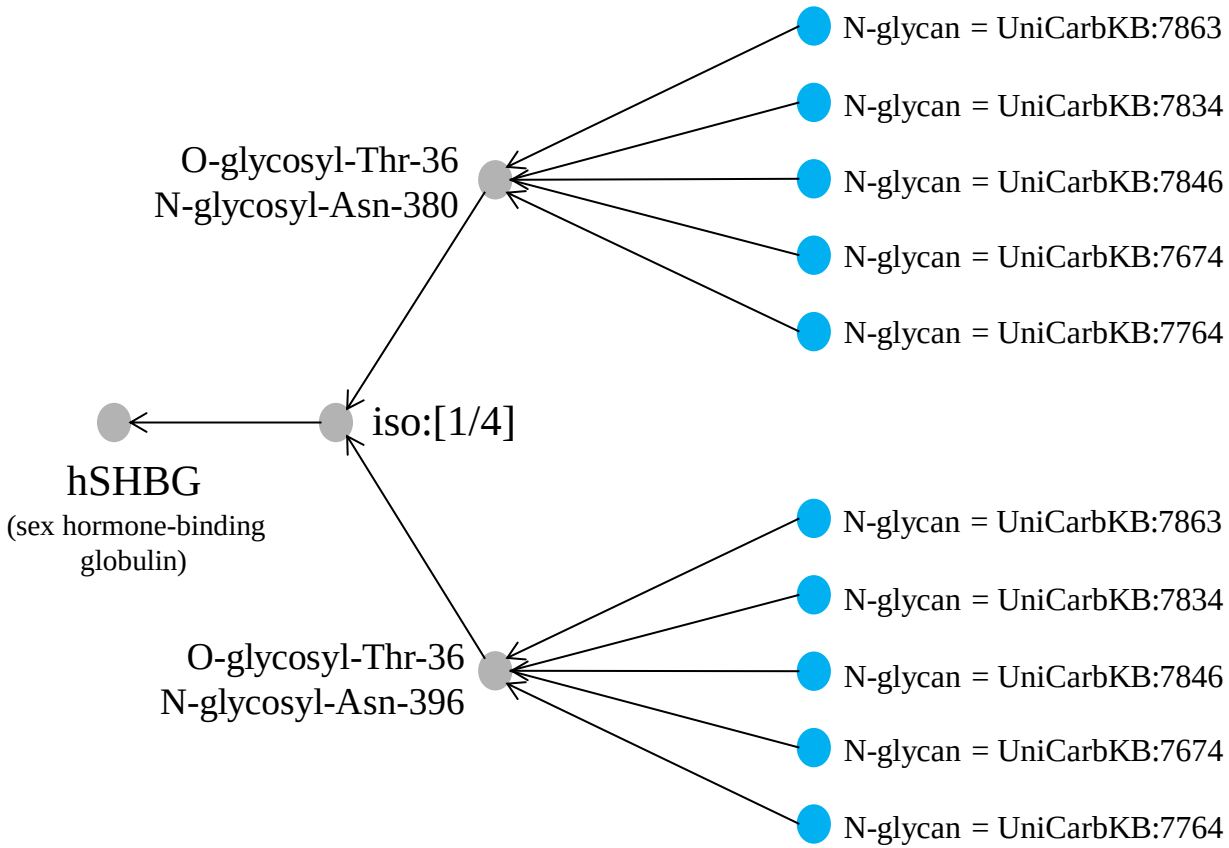
Leveraging Information: Predicting New Proteoforms



rat MAPK14 *1to1_ortholog_of* human MAPK14 (imported info)

rat MAPK14 *has_site* MKK kinase motif (imported info)

Future Directions: Glycobiology



“Oh the pain, the pain of it all”

- Reaching consensus
- Make it right, or make it useful
- Improvement vs content completeness

The PRO Consortium

Collaborating Institutions:

Principle Investigators:

Protein Information Resource (PIR)	Cathy Wu, Director of PIR
The Jackson Laboratory	Judith Blake, Associate Professor, Jackson Laboratory
NYS Center of Excellence in Bioinformatics and Life Sciences	Barry Smith, Professor, University at Buffalo
Reactome	Peter D'Eustachio, Associate Professor, NYU Med Ctr

Funding & Sponsors:

PRO is funded by NIGMS / NIH grants 1R01GM080646-01, 3R01GM080646-04S1, 3R01GM080646-04S2, 5R01GM080646-05