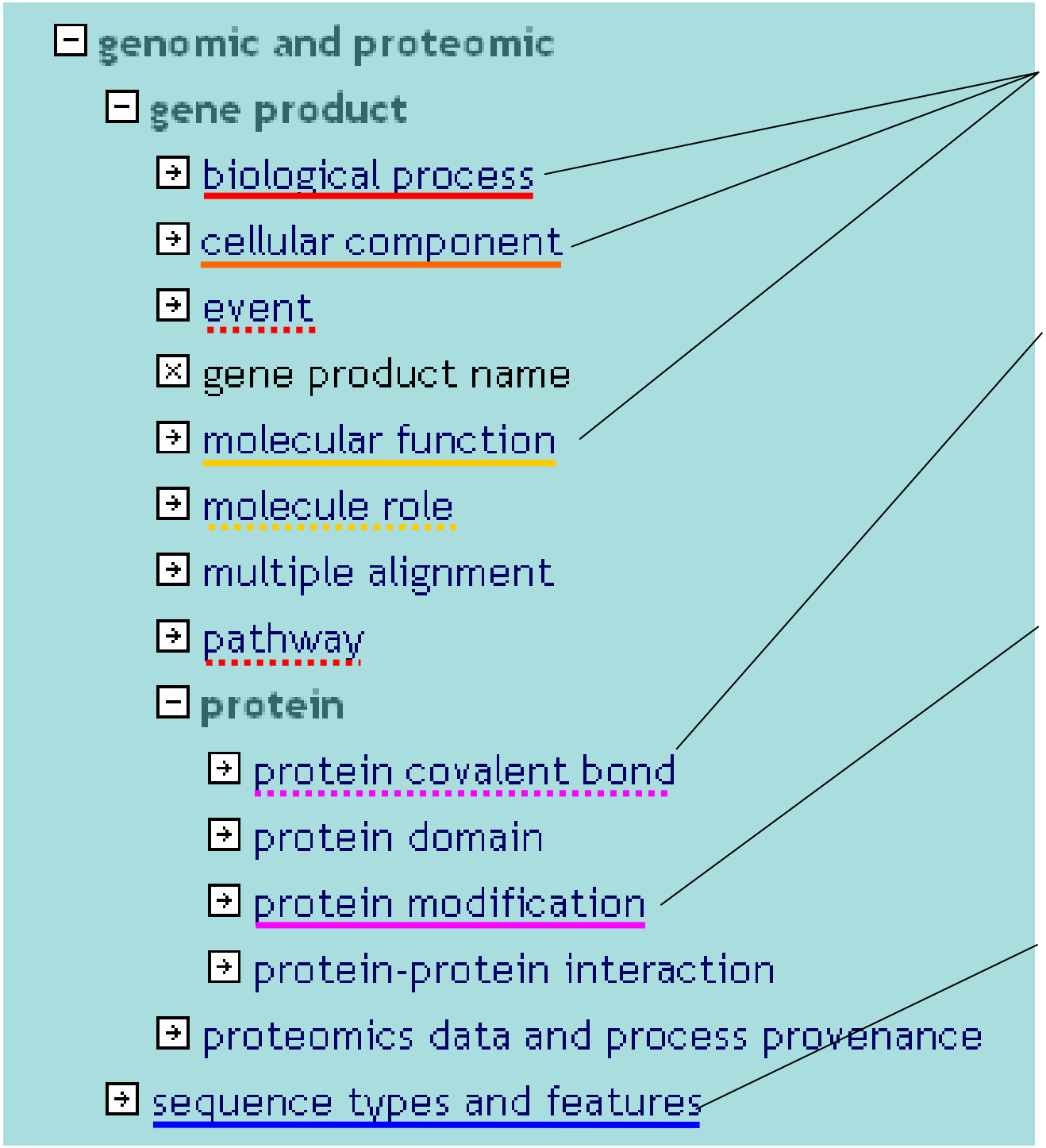
Framework for a Protein Ontology

TMBIO
November 2006

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



GO: ontologies that pertain, in part, to the locations, the processes, and the functions of proteins

PSI-MOD: ontology that describe the possible modifications to protein amino acid residues

SO: ontology that can describe the possible causes of protein sequence, expression, or structure changes

DO: ontology that can describe the possible effects of protein sequence, expression, or structure changes

Mothers against decapentaplegic homolog 2

Smad 2

GO annotation of SMAD2_HUMAN:

Cellular Component:

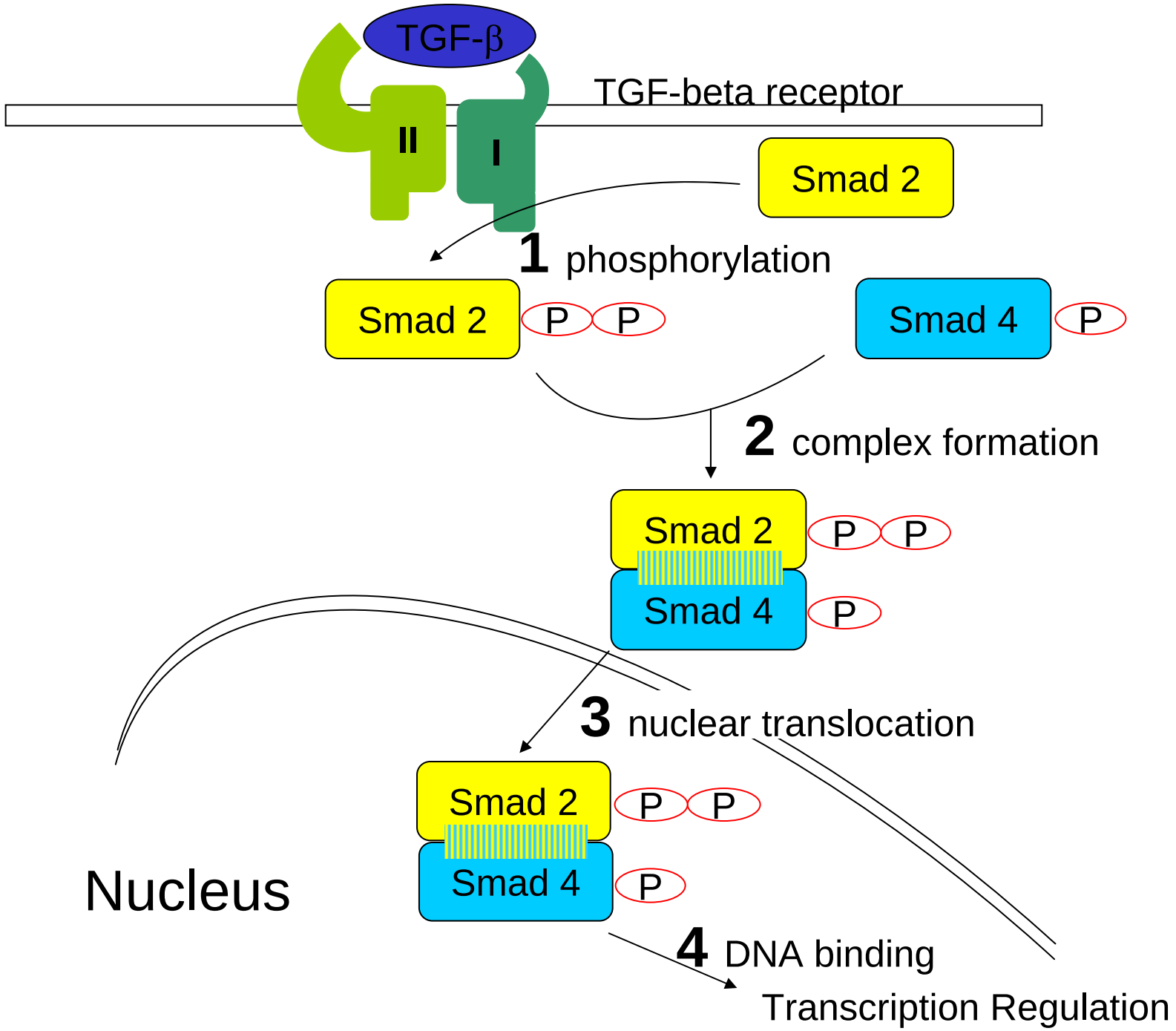
- nucleus

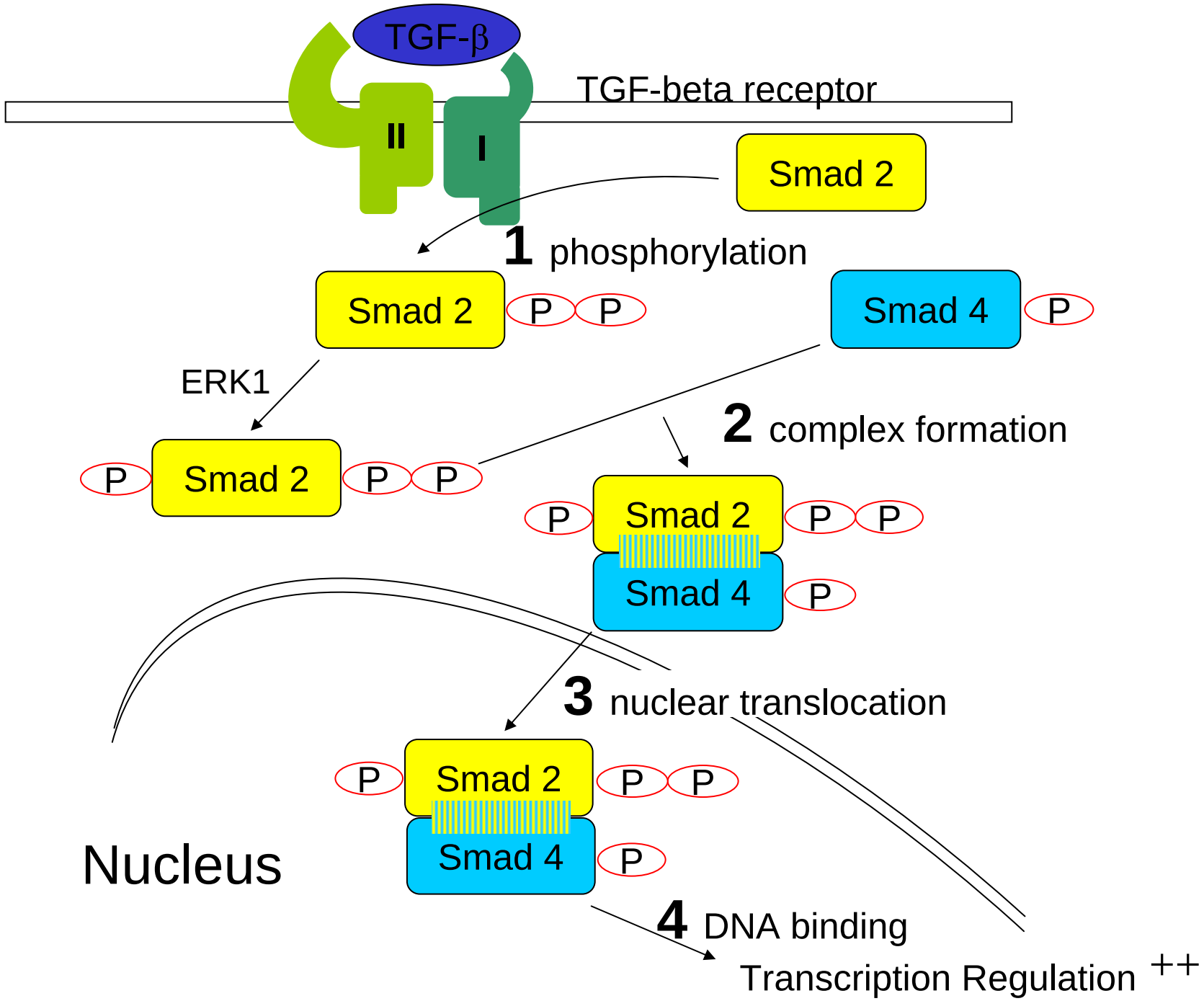
Molecular Function:

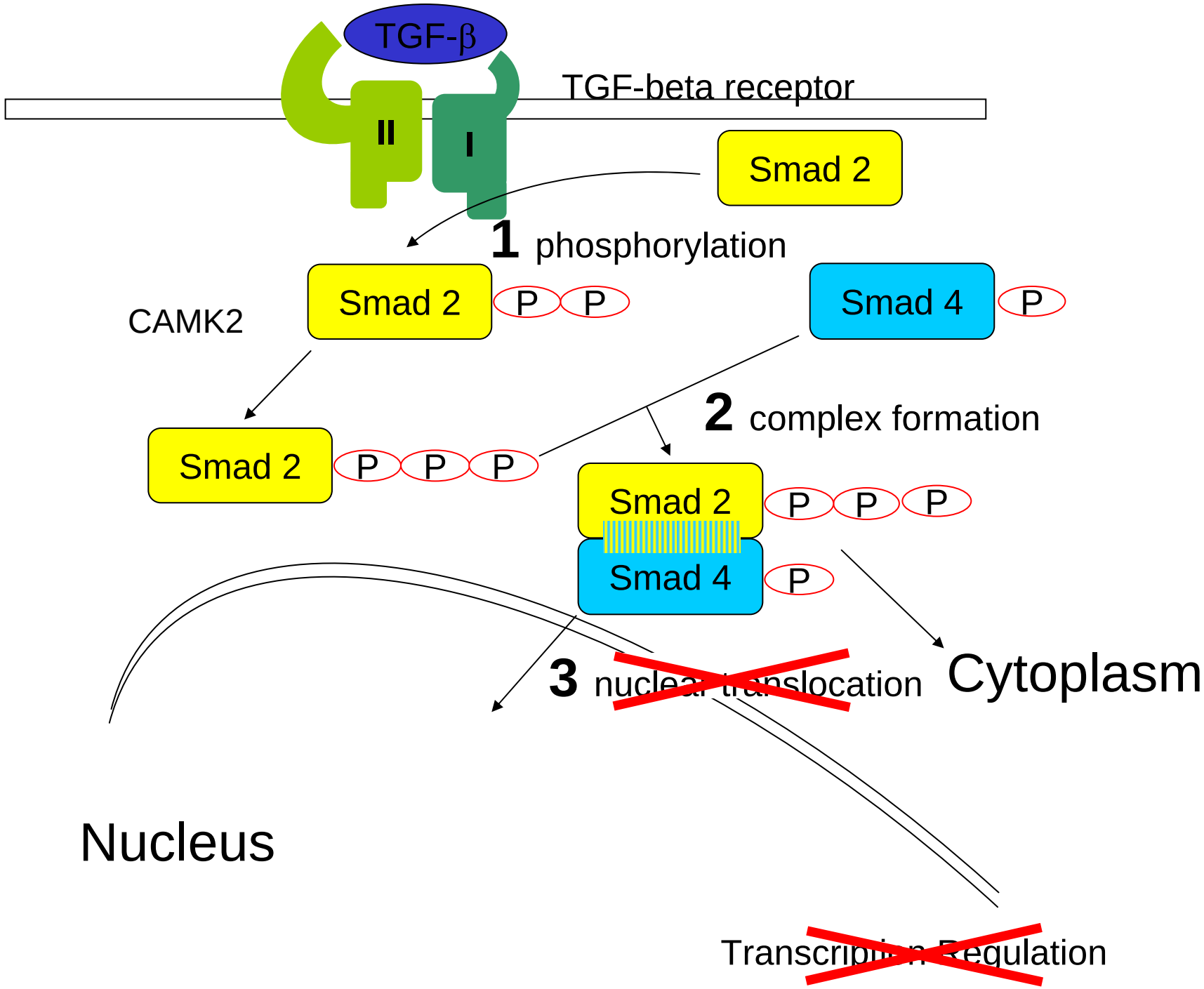
- protein binding

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
Smad 2
P
P
Smad 4
P
Nucleus
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
P
Smad 2
P P
Smad 4
P
Nucleus
4 DNA binding
Transcription Regulation ++


TGF-β
TGF-beta receptor
II
I
Smad 2
1 phosphorylation
CAMK2
Smad 2
P P
Smad 4
P
2 complex formation
Smad 2
P P P
Smad 2
P P P
Smad 4
P
3 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

Important Considerations

- Need to consider the various forms a protein might take
- Need to provide connections to established ontologies
- Need to account for the possibility that a protein might not share the traits of its parent or siblings

%**PRO:00000010** Smad2

<**PRO:00000011** Smad2 sequence 1 (long form)

>**PRO:00000012** Smad2 sequence 1 phosphorylated form

%**PRO:00000013** Smad2 sequence 1, TGF-β receptor I-phosphorylated

%**PRO:00000014** Smad2 sequence 1, TGF-β receptor I and ERK1-phosphorylated

has_modification **MOD:**O-phosphorylated L-serine

has_modification **MOD:**O-phosphorylated L-threonine

has_function **GO:** TGF-β receptor, pathway-specific cytoplasmic mediator activity

has_function **GO:**SMAD binding

has_function **GO:**transcription coactivator activity

participates_in **GO:**signal transduction

participates_in **GO:**SMAD protein heteromerization

participates_in **GO:**regulation of transcription, DNA-dependent

located_in **GO:**nucleus

part_of **GO:**transcription factor complex

%**PRO:00000015** Smad2 sequence 1, TGF-β receptor I and CAMK2-phosphorylated

<**PRO:00000016** Smad2 sequence 2 (short form) - splice variant

>**PRO:00000017** Smad2 sequence 2 phosphorylated form

%**PRO:00000018** Smad2 sequence 2, TGF-β receptor I-phosphorylated

<**PRO:00000019** Smad2 sequence 3 - genetic variant related to colorectal carcinoma

has_agent **SO:** amino_acid_substitution

lacks_modification **MOD:** phosphorylated residue

lacks_function **GO:** transcription coactivator activity

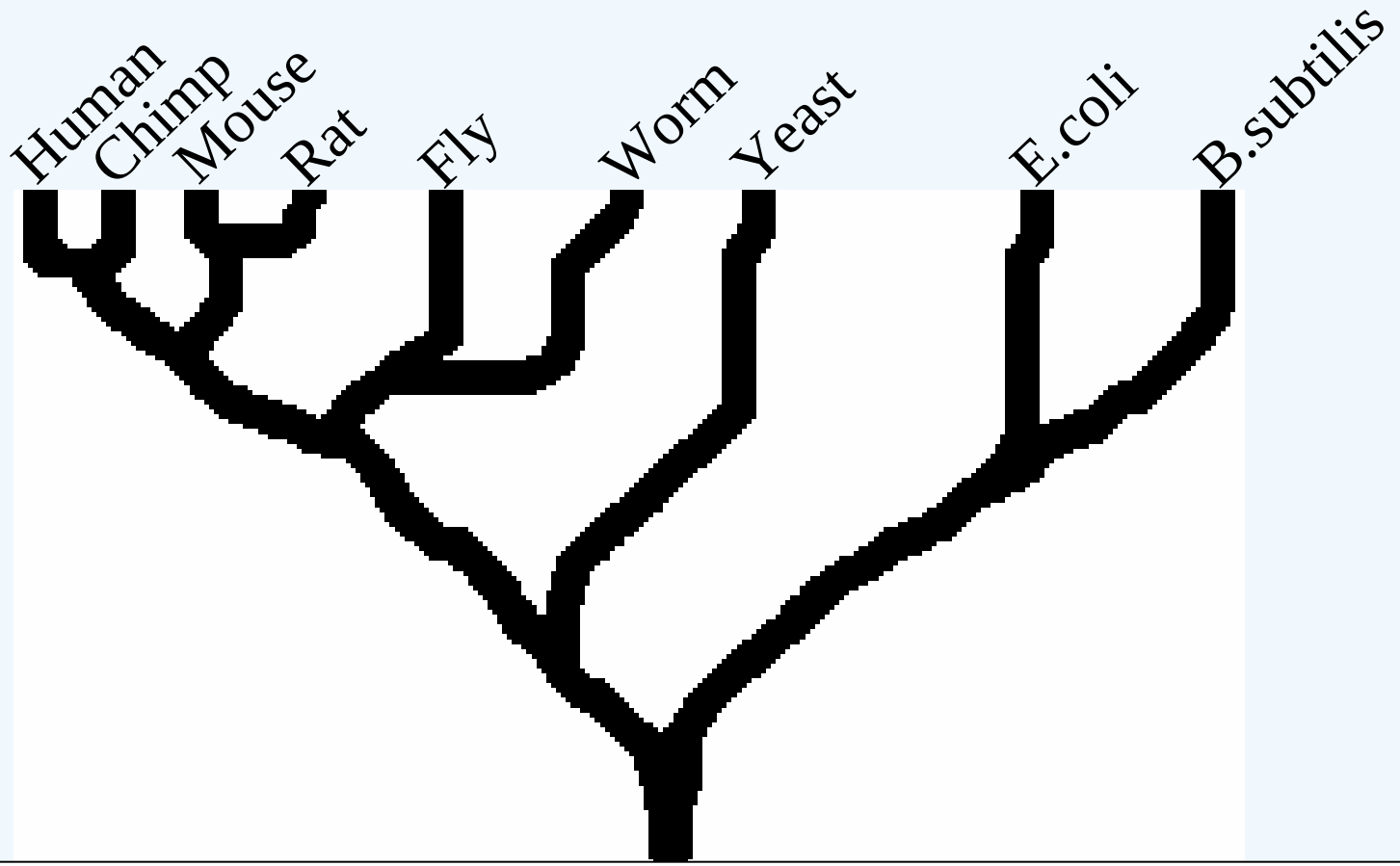
agent_of **DO:** carcinoma of the large intestine

%	is_a
<	variant_of
>	derives_from

Important Considerations

- Need to consider the various forms a protein might take
- Need to provide connections to established ontologies
- Need to account for the possibility that a protein might not share the traits of its parent or siblings
- Need to take advantage of model organism data to generate hypotheses about human biology

Implications of Protein Evolution

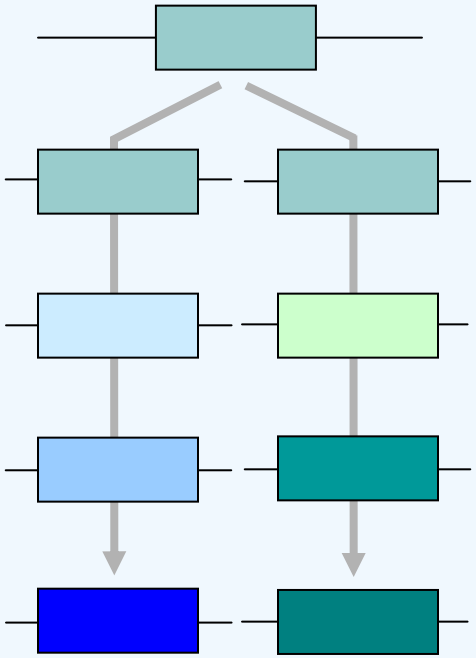


- Conclusions from experiments performed on proteins from one organism are often applicable to the homologous protein from another organism.

- Information learned about existing proteins allows us to infer the properties of ancestral proteins.

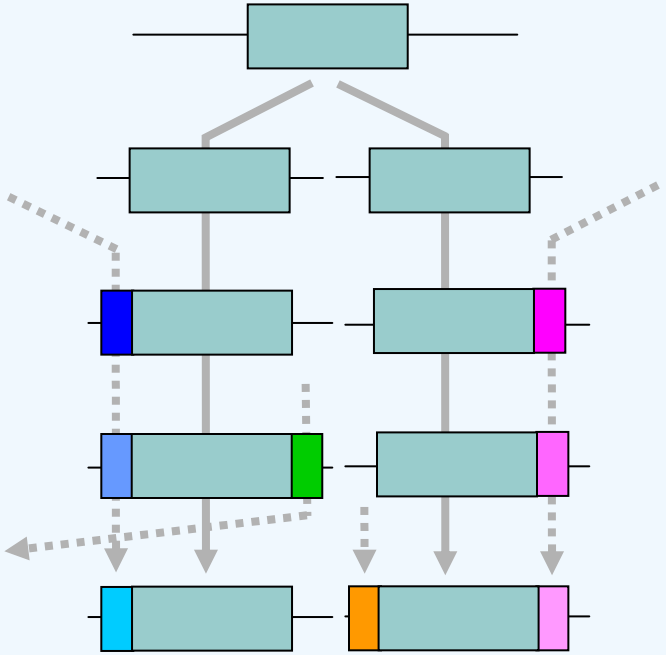
Protein Evolution

Sequence changes



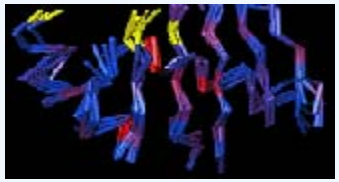
With enough similarity, one can trace back to a common origin

Domain shuffling



What about these?

Levels of Protein Classification

Evolutionary Divergence	*End-to-End Alignment*	*Example*	*Conservation*	*Database*
Very Ancient	**Domain structure**		**Very general biochemical activity**	**SCOP Superfamily**

Ancient	**Domain sequence**	`CVSGSGNNTSITATGGVVDLQSSSAVKVRSTK`
		`CYKSG---IQVRLGEDNINVVEGNEQFISASK`
		`* ... .: . ::: ... : ::*`
			General biochemical activity	**Pfam**

Recent	**Protein sequence** **Domain architecture**		**Specific biochemical activity**	**PIRSF**

Functionally Specialized	**Protein sequence**
	Domain architecture
		`CYKSRIQVRLGEHNIDVLEGNEQFINAAKIIT`
		`CYKSGIQVRLGEDNINVVEGNEQFISASKSIV`
		`**** *******.**:*:*******.*:* *.`
			Specific biological function	**PANTHER**

TGM3 & EPB42

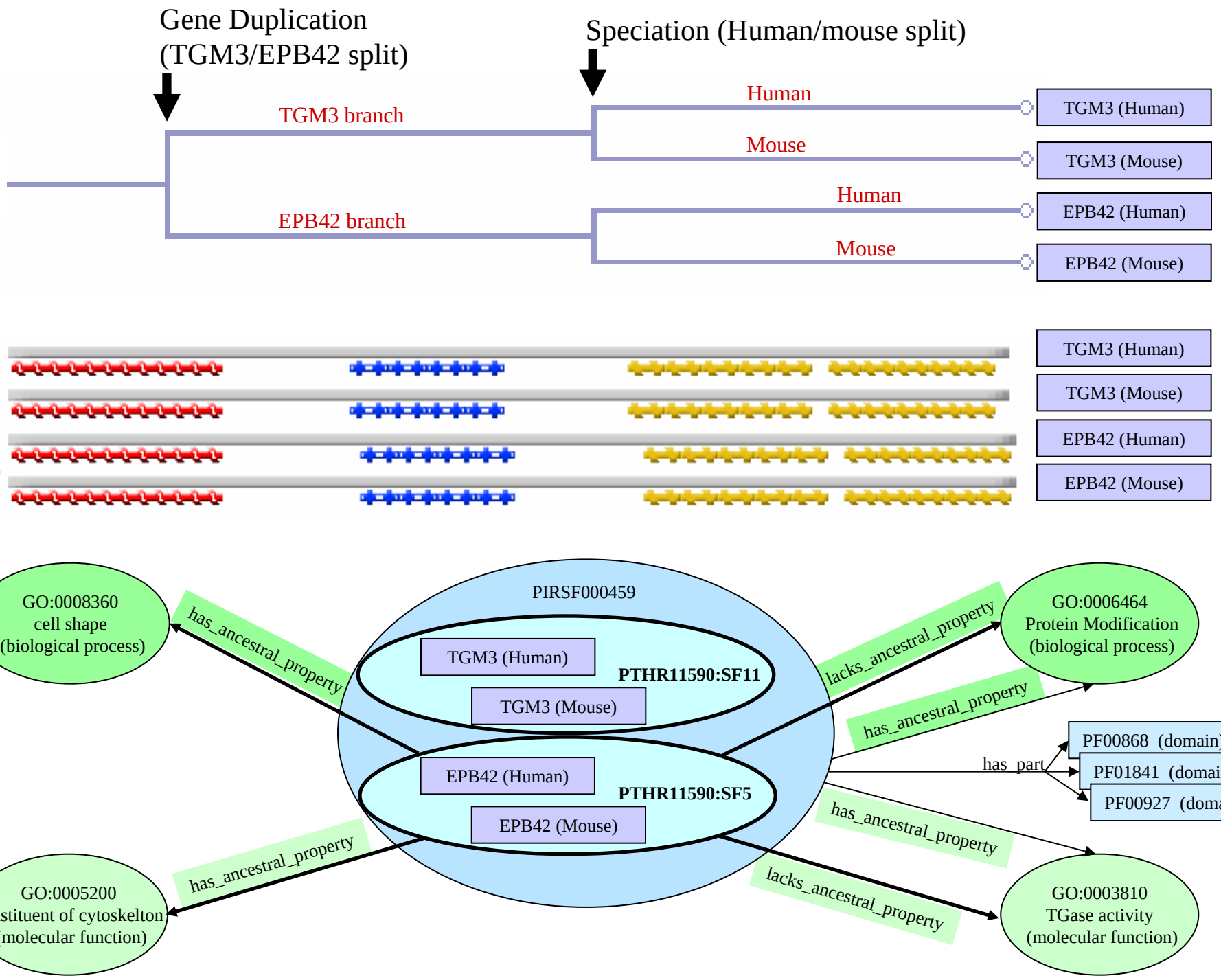
TGM3 = Protein-glutamine gamma-glutamyltransferase
(Transglutaminase; involved in protein modification)

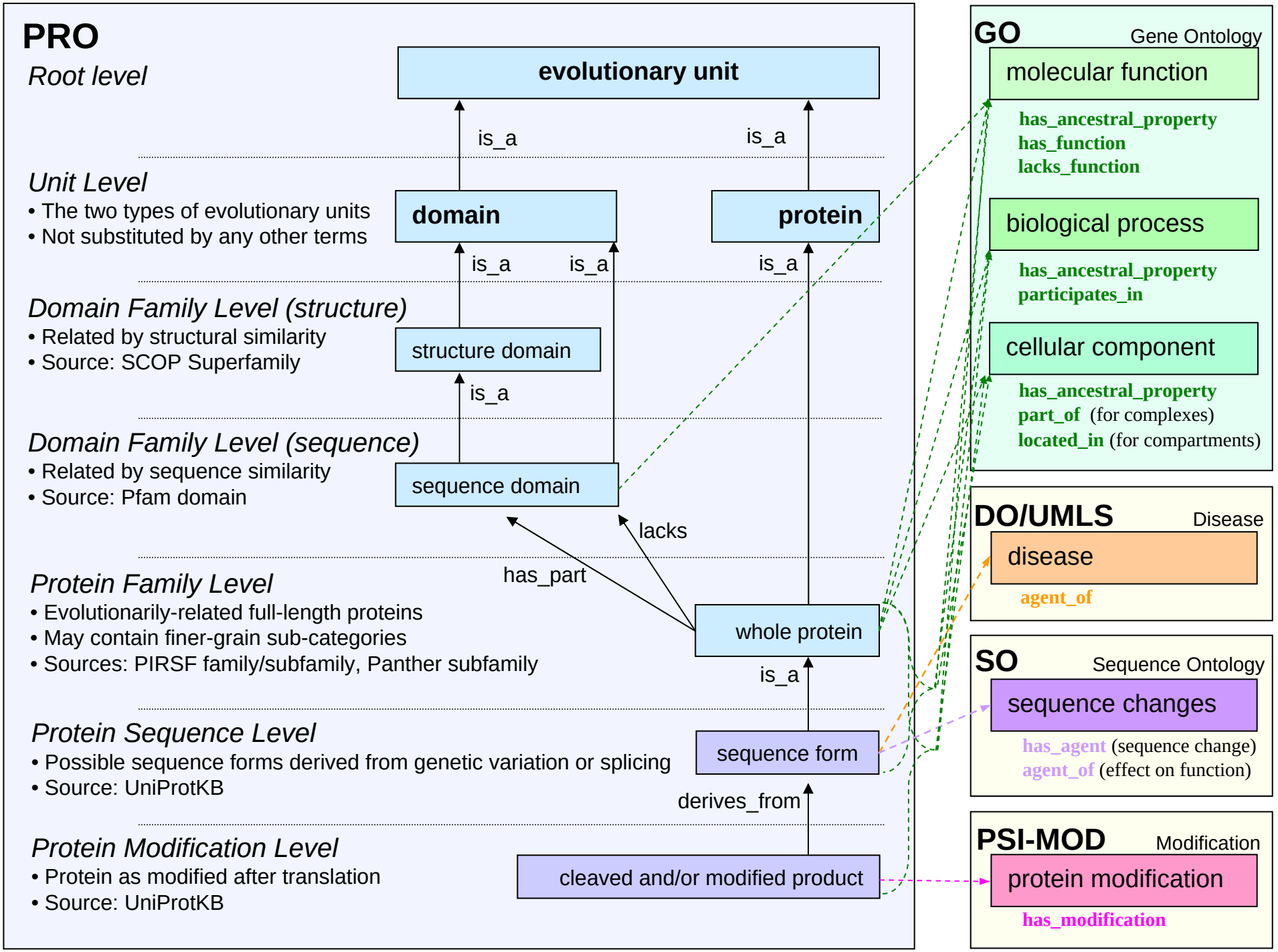
EBP42 = Erythrocyte membrane protein band 4.2
(Constituent of cytoskeleton; involved in cell shape)

Q: Are these related?

Q: What is known?

Q: How to capture?


Gene Duplication
(TGM3/EPB42 split)
Speciation (Human/mouse split)
TGM3 branch
EPB42 branch
Human
Mouse
Human
Mouse
TGM3 (Human)
TGM3 (Mouse)
EPB42 (Human)
EPB42 (Mouse)
TGM3 (Human)
TGM3 (Mouse)
EPB42 (Human)
EPB42 (Mouse)
PIRSF000459
PTHR11590:SF11
PTHR11590:SF5
TGM3 (Human)
TGM3 (Mouse)
EPB42 (Human)
EPB42 (Mouse)
GO:0008360
cell shape
(biological process)
GO:0005200
constituent of cytoskelton
(molecular function)
GO:0006464
Protein Modification
(biological process)
GO:0003810
TGase activity
(molecular function)
has_ancestral_property
has_ancestral_property
lacks_ancestral_property
has_ancestral_property
has_ancestral_property
lacks_ancestral_property
has_part
PF00868 (domain)
PF01841 (domain)
PF00927 (domain)


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
lacks
has_part
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

PRO Team (so far…)

- **Principle Investigators**
  - Cathy Wu (PIR at GUMC)
  - Judith Blake (The Jackson Laboratory)
  - Hongfang Liu (GUMC)
  - Barry Smith (SUNY Buffalo)

- **Curators**
  - Darren Natale (PIR at GUMC)
  - Cecilia Arighi (PIR at GUMC)
  - Winona Barker (PIR at GUMC)
  - Zhang-zhi Hu (PIR at GUMC)