**PIR**
A **UniProt** CONSORTIUM MEMBER
Protein Information Resource

```
TLALPN----RKAVADHLLM
LIGCLRNCSAVTAAAKQLAE
VTGFSN----AKTTAQHVKK
.:.*:.
```

# Protein Information Resource

**Integrated Protein Informatics Resource for Genomic , Proteomic and Systems Biology Research**

For over four decades the Protein Information Resource (PIR) has provided databases and protein sequence analysis tools to the scientific community, including the Protein Sequence Database, which grew out from the Atlas of Protein Sequence and Structure, edited by Margaret Dayhoff [1965-1978]. Currently, PIR major activities include: i) **UniProt** (Universal Protein Resource) development, ii) **iProClass** protein data integration and ID mapping, iii) **PRO** protein ontology, and iv) **iProLINK** protein literature mining and ontology development.
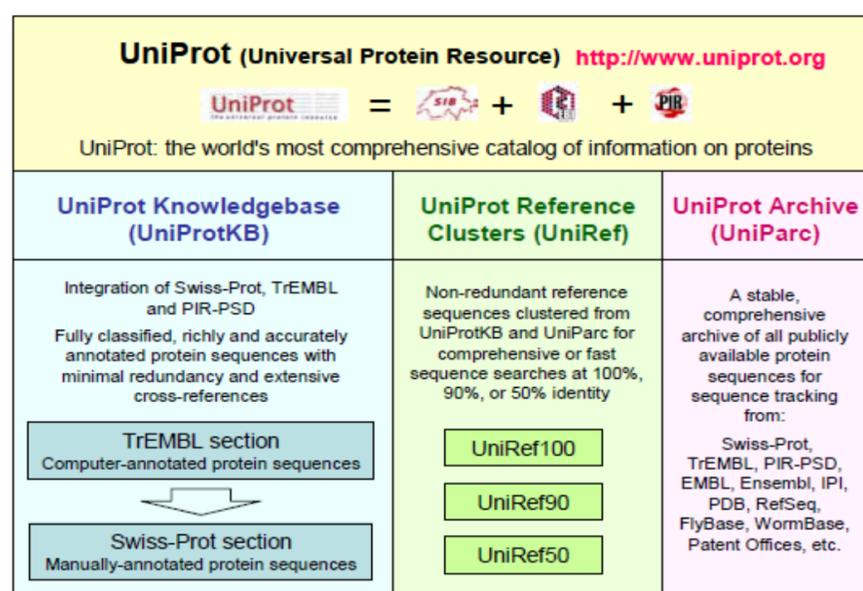
## UniProt – Universal Protein Resource

**UniProt**

### What is UniProt?
UniProt is the central resource for storing and interconnecting information from large and disparate sources and the most comprehensive catalog of protein sequence and functional annotation.

### When to use UniProt Databases
Use UniProtKB to retrieve curated, reliable, comprehensive information on proteins.
Use UniRef to decrease redundancy and speed up sequence similarity searches.
Use UniParc to access to archived sequences and their source databases.



**UniProt** (Universal Protein Resource)  http://www.uniprot.org

**UniProt** = SIB + EBI + PIR

UniProt: the world's most comprehensive catalog of information on proteins

| UniProt Knowledgebase (UniProtKB) | UniProt Reference Clusters (UniRef) | UniProt Archive (UniParc) |
|---|---|---|
| Integration of Swiss-Prot, TrEMBL and PIR-PSD. Fully classified, richly and accurately annotated protein sequences with minimal redundancy and extensive cross-references | Non-redundant reference sequences clustered from UniProtKB and UniParc for comprehensive or fast sequence searches at 100%, 90%, or 50% identity | A stable, comprehensive archive of all publicly available protein sequences for sequence tracking from: Swiss-Prot, TrEMBL, PIR-PSD, EMBL, Ensembl, IPI, PDB, RefSeq, FlyBase, WormBase, Patent Offices, etc. |
| TrEMBL section Computer-annotated protein sequences | UniRef100 | |
| Swiss-Prot section Manually-annotated protein sequences | UniRef90 | |
| | UniRef50 | |

## iProClass – Integrated Protein Knowledgebase

**iProClass**
Integrated Protein Knowledgebase
iPC

### What is iProClass?
iProClass provides extensive data integration of over 90 biological databases, with protein ID mapping service, and executive summary descriptions of proteins for UniProtKB and selected UniParc  protein sequences.

### When to use iProClass
Use iProClass to retrieve comprehensive, upto-date information about a protein, including function,  pathway, interactions, family classification,  structure and structural classification, gene and genomes, ontology, literature, and taxonomy. Use iProClass to access to ID mapping, protein BioThesaurus and related sequences.
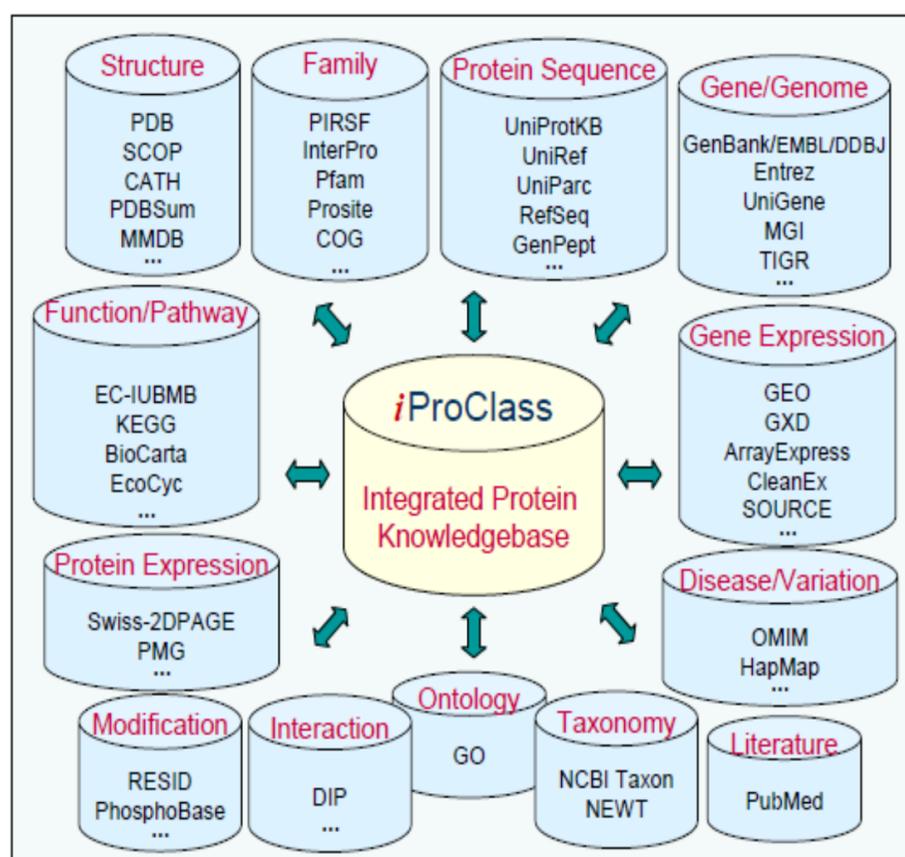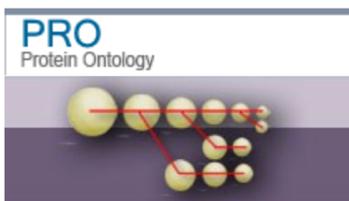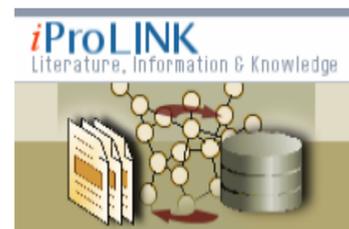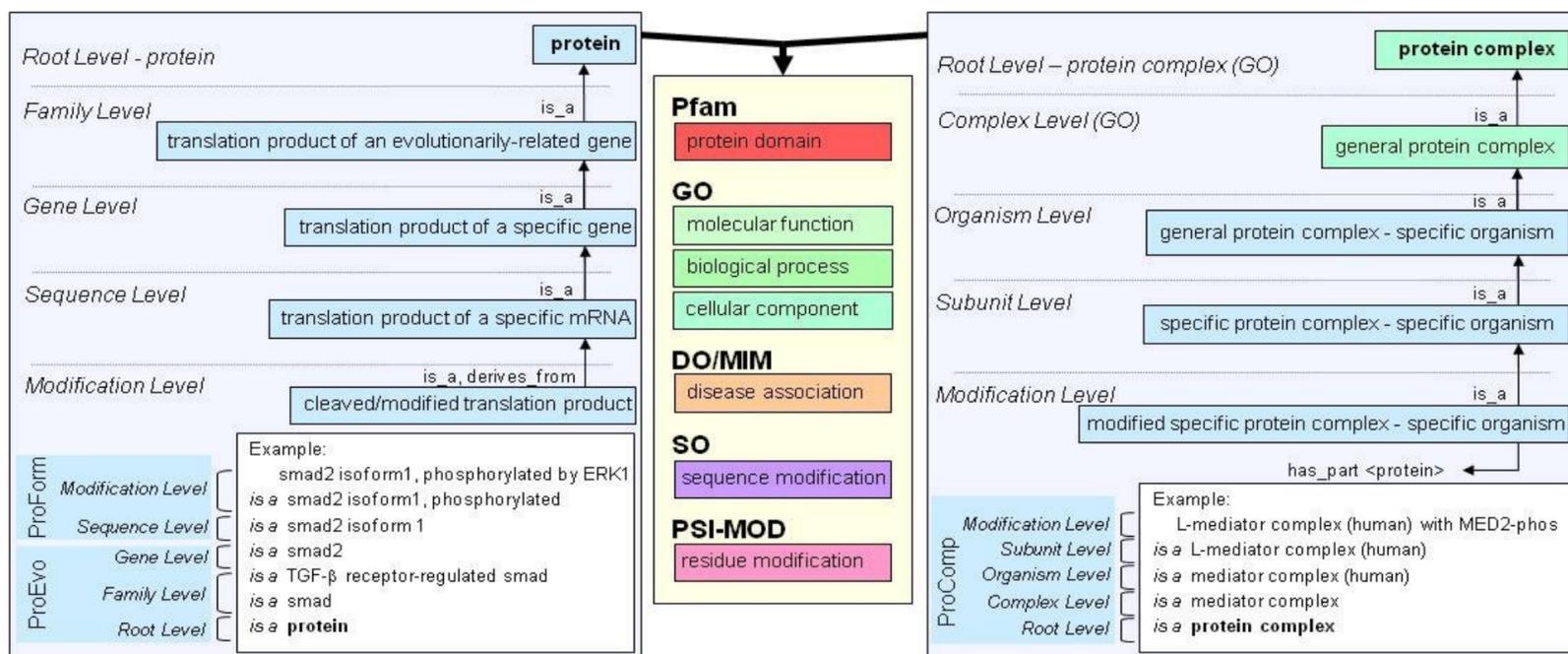
# PRO– Protein Ontology

## What is PRO?
PRO is a formal representation of protein objects, providing both descriptions of these objects and the relationships between them.

## When to use PRO?
PRO encompasses a sub-ontology of proteins based on evolutionary relatedness (ProEvo) and a sub-ontology of the multiple protein forms produced from a given gene (ProForm). PRO is interoperable with other OBO Foundry ontologies--such as the Sequence Ontology (SO) and the Gene Ontology (GO)--that provide representations of protein qualities. This interoperability facilitates cross-species comparisons, pathway analysis, disease modeling, and the generation of new hypotheses through data integration and machine reasoning.
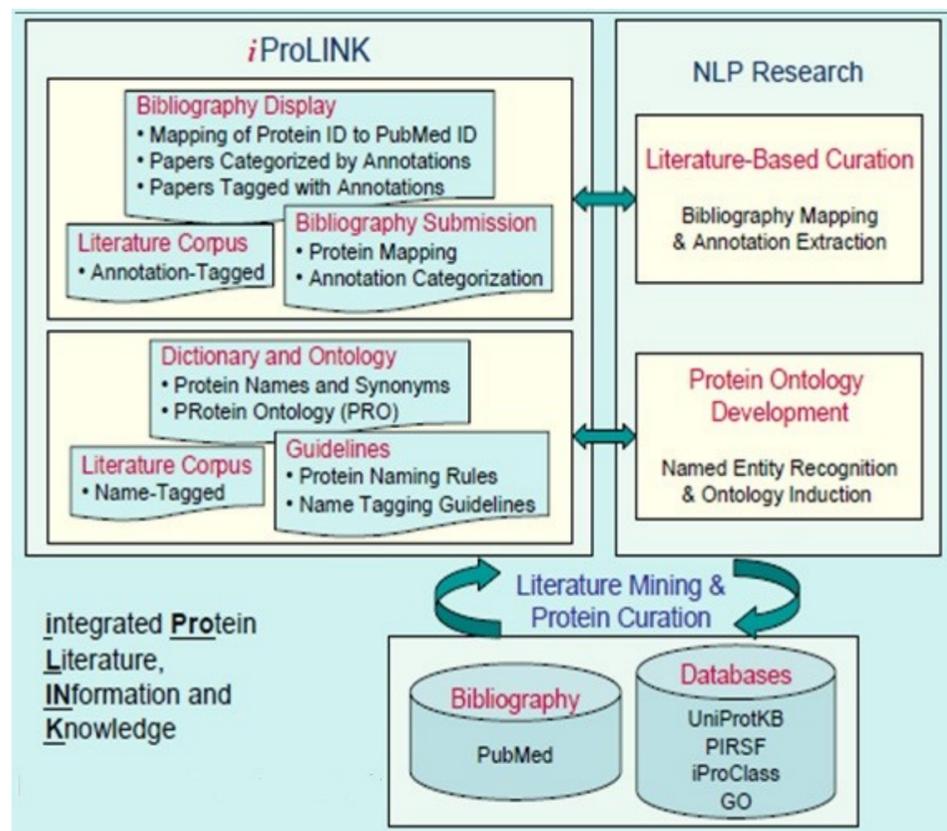
proteininformationresource.org/pro

### Left diagram (ProForm/ProEvo)

| Level | |
|---|---|
| Root Level - protein | protein |
| Family Level | is_a — translation product of an evolutionarily-related gene |
| Gene Level | is_a — translation product of a specific gene |
| Sequence Level | is_a — translation product of a specific mRNA |
| Modification Level | is_a, derives_from — cleaved/modified translation product |

**ProForm / ProEvo Example:**
- smad2 isoform1, phosphorylated by ERK1
- *is a* smad2 isoform1, phosphorylated — Modification Level
- *is a* smad2 isoform 1 — Sequence Level
- *is a* smad2 — Gene Level
- *is a* TGF-β receptor-regulated smad — Family Level
- *is a* smad
- *is a* **protein** — Root Level

### Middle boxes

**Pfam**
- protein domain

**GO**
- molecular function
- biological process
- cellular component

**DO/MIM**
- disease association

**SO**
- sequence modification

**PSI-MOD**
- residue modification

### Right diagram (ProComp)

| Level | |
|---|---|
| Root Level – protein complex (GO) | protein complex |
| Complex Level (GO) | is_a — general protein complex |
| Organism Level | is_a — general protein complex - specific organism |
| Subunit Level | is_a — specific protein complex - specific organism |
| Modification Level | is_a — modified specific protein complex - specific organism |
| | has_part <protein> |

**ProComp Example:**
- L-mediator complex (human) with MED2-phos
- *is a* L-mediator complex (human) — Modification Level
- *is a* mediator complex (human) — Subunit Level
- *is a* mediator complex — Organism Level
- *is a* mediator complex — Complex Level
- *is a* **protein complex** — Root Level

---

# iProLINK – Integrated Protein Literature, Information and Knowledge

## What is iProLINK?
iProLINK provides annotated literature, protein name dictionary, and other information to facilitate Natural Language Processing technology development in literature mining, database curation, protein name tagging and ontology.

## When to use iProLINK
Use iProLINK to obtain literature sources that describe protein entries (Bibliography Mapping), to map protein/gene names to UniProtKB entries (BioThesaurus), to obtain annotated data sets for developing text mining algorithms, to mine literature for protein phosphorylation (RLIMS-P).

proteininformationresource.org/iprolink



iProLINK

**Bibliography Display**
- Mapping of Protein ID to PubMed ID
- Papers Categorized by Annotations
- Papers Tagged with Annotations

**Literature Corpus**
- Annotation-Tagged

**Bibliography Submission**
- Protein Mapping
- Annotation Categorization

**Dictionary and Ontology**
- Protein Names and Synonyms
- PRotein Ontology (PRO)

**Literature Corpus**
- Name-Tagged

**Guidelines**
- Protein Naming Rules
- Name Tagging Guidelines

**NLP Research**

**Literature-Based Curation**
Bibliography Mapping & Annotation Extraction

**Protein Ontology Development**
Named Entity Recognition & Ontology Induction

integrated **Pro**tein **L**iterature, **IN**formation and **K**nowledge

**Literature Mining & Protein Curation**

**Bibliography** — PubMed

**Databases** — UniProtKB, PIRSF, iProClass, GO

---

Protein Information Resource

**PIR**