

PIRSF Protein Family Classification System

Anastasia Nikolskaya, Cecilia Arighi, Winona Barker, Hongzhan Huang, Raja Mazumder, Darren Natale, Sona Vasudevan, C.R. Vinayaka, Lai-Su Yeh, Cathy Wu

Protein Information Resource, Georgetown University Medical Center
ann2@georgetown.edu <http://pir.georgetown.edu/>

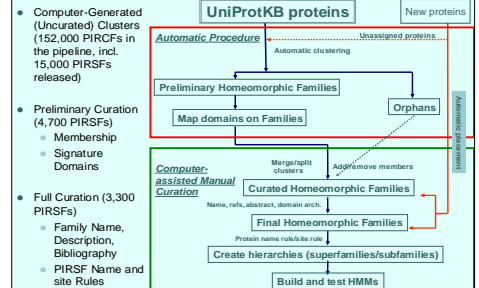
Abstract

The PIRSF protein classification system reflects evolutionary relationships of full-length proteins and domains. PIRSF families are extensively curated using a bioinformatics infrastructure implemented in a J2EE framework. Expert manual curation includes membership, annotation of specific biological functions, biochemical activities, and sequence features. Novel functional predictions for uncharacterized "hypothetical" proteins and protein families are routinely made in the annotation process. Fully curated families and their protein members provide basis for rich and accurate functional annotation of protein sequences in the UniProt Knowledgebase. The PIRSF database is accessible at <http://pir.georgetown.edu/pirsf/>

PIRSF Classification System

- PIRSF:**
 - Reflects **evolutionary relationships of full-length proteins**
 - A **network structure from superfamilies to subfamilies**
- Definitions:**
 - Homeomorphic Family:** Basic Unit
 - Homologous:** Common ancestry, inferred by sequence similarity
 - Homeomorphic:** Full-length similarity & common domain architecture
 - Hierarchy:** Flexible number of levels with varying degrees of sequence conservation
 - Network Structure:** allows multiple parents
- Advantages:**
 - Annotate both general biochemical and specific biological functions
 - Accurate propagation of annotation and development of standardized protein nomenclature and ontology

Creation and curation of PIRSFs



PIRSF Classification System

Pfam Domain	PIRSF Superfamily	PIRSF Homeomorphic Family	PIRSF Homeomorphic Subfamily
PF0735: Ku70/Ku80 beta-barrel domain	PIRSF00001: Ku DNA-binding complex Ku70/80 subunits	PIRSF00303: Ku70 subunit PIRSF01657: Ku80 subunit	
PF0219: Insulin-like growth factor binding protein (IGFBP)		PIRSF00198: IGFBP PIRSF01823: IGFBP-related protein, MAC25 type	PIRSF00001: IGFBP-1 PIRSF00006: IGFBP-6
PF01817: Chorismate mutase (CM)		PIRSF01731: CM of AroQ class, eukaryotic type PIRSF01501: CM of AroQ class, prokaryotic type PIRSF01500: Bifunctional CMPOT (P-protein) PIRSF01499: Bifunctional CMPDH (T-protein)	
PF02153: Prephenate dehydratase (PDH)		PIRSF01489: Bifunctional CMPDH (T-protein) PIRSF0786: PDH, feedback inhibition-insensitive PIRSF005547: PDH, feedback inhibition-sensitive	

PIRSF domain architecture display:
CM domain
PF01817

PIRSF Report

Family-driven Protein Annotation

Family-Driven Protein Annotation

Objective: Optimize for protein annotation

- PIRSF Classification Name**
 - Reflects the function when possible
 - Indicates the maximum specificity that still describes the entire group
 - Standardized format
 - Name tags: validated, tentative, predicted, functionally heterogeneous
- Hierarchy**
 - Subfamilies increase specificity (kinase -> sugar kinase -> hexokinase)
- Name Rules**
 - Define conditions under which names propagate to individual proteins
 - Enable further specificity based on taxonomy or motifs
 - Names adhere to Swiss-Prot conventions (though we may make suggestions for improvement)
- Site Rules**
 - Define conditions under which features propagate to individual proteins

PIR Name Rules

- Account for functional variations within one PIRSF, including:
 - Lack of active site residues necessary for enzymatic activity
 - Certain activities relevant only to one part of the taxonomic tree
 - Evolutionarily-related proteins whose biochemical activities are known to differ
- Monitor such variables to ensure accurate propagation**
- Propagate other properties that describe function: EC, GO terms, misnomer info, pathway
- Name Rule types:**
 - "Zero" Rule
 - Default rule (only condition is membership in the appropriate family)
 - Information is suitable for every member
 - "Higher-Order" Rule
 - Has requirements in addition to membership
 - Can have multiple rules that may or may not have mutually exclusive conditions

PIR Site Rules

- Position-Specific Site Features:**
 - active sites
 - binding sites
 - modified amino acids
- Current requirements:**
 - at least one PDB structure
 - experimental data on functional sites: CATRES database (Thornton)
- Rule Definition:**
 - Select template structure
 - Align PIRSF seed members with structural template
 - Edit MSA to retain conserved regions covering all site residues
 - Build Site HMM from concatenated conserved regions

PCS Curation Platform as a Means of Community Annotation

PCS: Search and Retrieval

Retrieve all proteins sharing a common domain

Graphical Analysis Tool Integration

- Curator-guided clustering
- Single-linkage clustering using BlastClust
- Fixed-length coverage enforces homeomorphy
- Iterative procedure allows tree view

DAG Viewer

Two unrelated chorismate mutase groups:
1. PIRSF005965 (PF07736 domain)
2. All other PIRSFs (PF01817 domain)

"Orphans": no family classification