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 ZEKE 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 superfamily to subfamily

Definitions:
- Homogeneous Family: Basic Unit
- Homologous: Common ancestry, inferred by sequence similarity
- Homogeneous: Full-length similarity & common domain architecture
- Hierarchy: Flexible number of levels with varying degrees of sequence similarity

Advantages:
- Accurate general biochemical and specific biological functions
- Accurate propagation of annotation and development of standardized protein nomenclature and nomenclature

PIRSF Report

PIRSF Domain Architecture Display: CM domain P001617

Family-driven Protein Annotation

PIRSF Classification Name: Reflection of the function when possible
- Indicates the maximum specificity that still describes the entire group
- Standardized format
- Names target: classification predicted, functionally homogeneous
- Hierarchy: Standardized base name = subfamily name = family name

Name Rules:
- Define condition under which names propagate to individual proteins
- Define conditions under which names propagate to individual proteins
- Define conditions under which names propagate to individual proteins

PIRSF Site Rules

Position-Specific Site Features:
- Account for functional variations within one PIRSF, including
  - Lack of active site residues necessary for enzymatic activity
  - Centers activity relevant only to one part of the biochemical tree
- Evolutionarily related proteins whose biochemical activities are known to differ
- Monitor such variables to ensure accurate propagation

Name Rule types:
- "Dear" Rule:
  - Delivers use (longer context in membership in the appropriate family)
- Information is suitable for every member
- "Higher-Order" Rule:
  - Incorporates use in addition to membership
  - Can have multiple rules that stop or may not have mutually exclusive conditions

PIRSF Curation Platform as a Means of Community Annotation

PCS: Search and Retrieval
- Removes all proteins sharing a common domain

Graphical Analysis Tool Integration
- Curation-guided clustering
- Single-linkage clustering using BlockClust
- Fast alignment coverage enhances homology
- Iterative procedure allows low view

DAG Viewer
- "DAGers" no family classification

Acknowledgements: Uniprot is supported by the National Institutes of Health, grant # 1 U01 HG02712-01