

## PIRSF PROTEIN FAMILY CLASSIFICATION SYSTEM AND PROTEIN SEQUENCE ANNOTATION

Anastasia N. Nikolskaya

Protein Information Resource, Georgetown University Medical Center, Washington, DC  
ann2@georgetown.edu

High-throughput genome projects have resulted in a rapid accumulation of predicted protein sequences; however, experimentally verified information on protein function lags far behind. The common approach to inferring function of uncharacterized proteins based on sequence similarity to annotated proteins often results in over-identification, under-identification, or even misannotation. To facilitate accurate, consistent and rich functional annotation of proteins, Protein Information Resource (PIR, <http://pir.georgetown.edu/>) employs a classification-driven rule-based automated annotation method supported by a bioinformatics framework that provides data integration and associative analysis.

Towards this goal, PIR has developed the SuperFamily (PIRSF, <http://pir.georgetown.edu/pirsf/>) classification system. This classification, based on the evolutionary relationships of whole proteins, allows annotation of both specific biological and generic biochemical functions. The system adopts a network structure for protein classification from superfamily to subfamily levels. The primary PIRSF classification unit is the *homeomorphic family* whose members are *homologous* (sharing common ancestry) and *homeomorphic* (sharing full-length sequence similarity with common domain architecture). The PIRSF database consists of two data sets: preliminary computer-generated clusters and curated families. Families are curated for name, membership, parent-child relationships, domain architecture, and optional description and bibliography.

When an experiment yields a sequence (or a set of sequences), it is often necessary to assess the function of this protein based on sequence alone. The quality of this assessment is often critical for interpreting experimental results and making hypothesis for future experiments. Searching a protein sequence against curated PIRSFs provides faster and more accurate results than a BLAST search against an uncurated protein database because it avoids the pitfalls such as numerous erroneous annotations in the databases, best hits based on a domain secondary to the protein function, spurious hits etc. PIRSF allows *going from sequence to function* and getting curated, reliable and enriched information. The integrative approach leads to novel prediction and functional inference for uncharacterized proteins, allows systematic detection of genome annotation errors, and provides sensible propagation and standardization of protein annotation.

PIR recently joined the European Bioinformatics Institute and Swiss Institute of Bioinformatics to establish UniProt (<http://www.pir.uniprot.org/>), an international resource of centralized, value-added protein knowledge that unifies PIR, Swiss-Prot, and TrEMBL databases. PIRSF is an integral part of the UniProt annotation pipeline.