PIR: A Comprehensive Resource for Functional Analysis of Protein Sequences and Families

Anastasia Nikolskaya, Lai-Su Yeh
Protein Information Resource, Georgetown University Medical Center, Washington, DC

PIR (http://pir.georgetown.edu/) provides a comprehensive resource for the bioinformatics analysis of protein sequences and families, with the main purpose of accurate functional analysis and annotation of UniProt sequences. At the PIR, functional annotation of proteins is based on a bioinformatics framework that couples PIRSF protein family classification with the iProClass integrated protein database. This allows associative analysis using information on protein sequence, structure, function, and other systems biology information. The integrative approach has led to novel predictions and functional inference for previously uncharacterized proteins, to detection and correction of genome annotation errors, as well as to enhanced understanding of structure, function, and evolutionary relationships.

The PIRSF protein classification system (http://pir.georgetown.edu/pirsf/) is based on evolutionary relationships of full-length proteins and domains, has a network structure and has been developed to facilitate the propagation and standardization of protein annotation. The multiple levels of sequence diversity, from superfamilies to subfamilies, reflect different degrees of functional granularity.

The PIRSF database is central to the functional annotation of proteins in UniProt (Universal Protein Resource), both in the UniProt/TrEMBL (automatic annotation) and in UniProt/Swiss-Prot (expert manual annotation) sections of the UniProt Knowledgebase (UniProtKB).

The new PIRSF curation platform connects a set of analysis and visualization tools and a DAG editor to maximize throughput and minimize routine, error-prone tasks, thereby allowing scientists to provide richly and accurately curated network of protein families. Integrated tools include PIRClust (iterative BlastClust) for sequence clustering, CLustalW for multiple sequence alignment with neighbor-joining phylogenetic tree, a PIR taxonomy tree browser, and the SEED program for genome context and subsystem analysis.

Complementing the PIRSF family classification system, the data integration in the iProClass database facilitates functional exploration and comparative analysis of proteins. iProClass (http://pir.georgetown.edu/iproclass/) provides value-added descriptions of all UniProt proteins, with rich links to over 90 databases of protein family, function, pathway, interaction, modification, structure, genome, expression, ontology, literature, and taxonomy.

The integrative approach allows associative studies of protein family, domain, function, and structure using information on protein sequence, structure, function, and other system biology information. This includes drawing on various types of available information to provide a comprehensive picture that can lead to novel prediction and functional inference for previously uncharacterized proteins beyond sequence homology.

Using PIR resources to maximize the information about a given sequence will be demonstrated with case studies. This will include: (1) text and sequence search of the iProClass and PIRSF databases, (2) conserved motif search and analysis, (3) information retrieval from fully curated PIRSF families, (4) multiple sequence alignment, and (5) PIRClust (iterative BlastClust).