PIR-Class: An Object-Relational Protein Classification Database for Sequence Annotation and Genome Research

Cathy H. Wu, Chunlin Xiao, Zhenglin Hou, and Winona C. Barker
Protein Information Resource, National Biomedical Research Foundation,
Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, DC 20007-2195
wuc (xiaoc, houz, barker)@nbrf.georgetown.edu

The molecular sequence data continue to grow at an accelerating pace due to the Human Genome Project and other large sequencing projects. Advanced databases are needed to facilitate the retrieval of relevant information from the voluminous data and to provide insight into protein structure and function. Protein family classification is now well recognized as an effective approach for large-scale genomic sequence annotation and database organization.

PIR-Class is an object-relational protein database newly created by the Protein Information Resource1 to provide an integrated platform for describing comprehensive family relationships and structural and functional features of proteins. It is a secondary, value-added family database built upon the PIR-International Protein Sequence Database[1] and several other databases, including the ProClass protein family database[2] and the PIR-ALN family alignment database[3]. To support a complete database organization, the PIR superfamily/family serves as the underlying classification scheme for complete and non-overlapping clustering of all proteins. The classification at the whole protein level is directly associated with domains and motifs, which represent structural and functional building blocks.

The basic family entity has three major attributes: membership, annotation, and relationship. Membership lists all sequence entries, with summary information such as length distribution and taxonomic range. Functional annotation includes keywords, functional sites, and enzyme classification (EC), with links to metabolic pathway and enzyme databases such as KEGG and BRENDA. Structural annotation includes features such as structural motifs/sites, signal peptides, and transmembrane domains, with links to most closely related PDB structures and structural class databases such as SCOP and CATH. The family relationship integrates classification at both the global (full-length) and local (domain and motif) levels, with cross-references to other family databases, including Pfam, PRINTS, BLOCKS, and PROSITE.

Each family record in the classification database thus documents family relationships and features more comprehensively than any other single information resource. The classification database is being implemented in the Oracle object-relational database management system and will soon be available for on-line search from our WWW site[4] at http://pir.georgetown.edu/pirwww/search/pirclass.html. It will allow users to retrieve conveniently displayed family summaries for given sequences and to query for sequences and families with selected properties. The database facilitates classification-driven

annotation for protein sequence databases and complete genomes, as well as supporting structural and functional genomics research.

1 Barker *et al*., (2000) The Protein Information Resource (PIR), *Nucleic Acids Res*., 28: 41-44.
2 Huang *et al*., (2000) ProClass protein family database. *Nucleic Acids Res*., 28: 273-276.
3 Srinivasarao *et al*., (1999) PIR-ALN: A database of protein sequence alignments, *Bioinformatics*, 15: 382-390.
4 McGarvey *et al*., (2000) The PIR Web site: New resource for bioinformatics, *Bioinformatics*, 16: 1-3.

Back to Publications Page