PIR and *i*ProClass for Functional Genomics/Proteomics

Barker, W.C., Castro-Alvear, J., Chen, Y., Hu, Z., Huang, H., Ledley, R.S., Lewis, K.C., Orcutt, B.C., Suzek, B., Vinayaka, C.R., Wu, C.H., Yeh, L.L., and Zhang, J.
Protein Information Resource, National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, DC 20007-2195

The human genome project has revolutionized the practice of biology and the future potential of medicine. With the tremendous volume of genomic and molecular data being generated, computational approaches have become increasingly important for deriving and evaluating hypotheses. Protein family classification is an effective means for large-scale functional characterization of genes. The Protein Information Resource (PIR) is an integrated public resource for protein knowledge management using the family classification approach. The PIR Web site provides many protein databases, including the PIR-International Protein Sequence Database, and data analysis and mining tools, including search engines that combine searches of sequence similarity and database annotation to facilitate the analysis and functional identification of proteins. The iProClass database provides comprehensive family relationships at global (whole protein) and local (domain and motif/site) levels, as well as structural/functional classifications and features of proteins. The PIR superfamily/family organization allows non-overlapping clustering of all proteins. The iProClass currently consists of about 266,000 non-redundant PIR and SwissProt proteins organized with more than 30,000 superfamilies, 100,000 families, 3000 domains, 1300 motifs, 280 post-translational modification sites, and links to over 40 databases of protein families, structures, functions, genes, genomes, literature, and taxonomy. Future releases will be based on a new PIR non-redundant reference sequence database (NREF) containing more than 800,000 protein sequences. Protein and superfamily summary reports provide annotations such as membership information with length, taxonomy, and keyword statistics, extensive cross-references, and graphical display of domain and motif regions. The iProClass employs a modular structure for scalability and extendibility, thereby providing a framework for integration of new data and/or software components in a distributed networking environment. The PIR databases are implemented in Oracle 8i object-relational database system and freely accessible from our web site at http://pir.georgetown.edu/.
Supported by NLM grant LM05798 and NSF grant DBI-9974886

Presented at the 41st Annual Meeting for the American Society for Cell Biology, Washington, D.C., 2001

Back to Publications Page