Superfamily System for Functional Classification of Protein Sequences: Recognizing Cellular Phosphoesterases in Sequenced Genomes

W. C. Barker, H. Huang, L.S. Yeh, and C.H. Wu
Protein Information Resource, National Biomedical Research Foundation,
Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, DC
20007-2195

Classification of protein sequences is very important for large-scale functional characterization of genes. For accurate identification, it is necessary to classify proteins both by domain and motif identification and on the basis of end-to-end similarity and domain architecture. The protein superfamily organization of the PIR-International Protein Sequence Database (PIR-PSD) is the only comprehensive protein classification system that is based on global similarity and identical domain arrangement. We have developed an integrated system that includes automated procedures and Web interfaces for rapid and accurate classification of large numbers of protein sequences into comprehensive and non-overlapping families and superfamilies. These procedures are acilitated by the PIR Annotation and Similarity Database, which includes pre-computed FASTA neighbors of PIR-PSD entries, and the PIR HMM Homology Domain Database, which contains homology domain information for PIR-PSD entries. Superfamilies are formed based on overall similarity and identical domain arrangement. Currently, we have classified about 70% of 240,000 PIR-PSD sequences into 33,000 superfamilies.
We illustrate how using motif, domain, and superfamily classification allows previously uncharacterized genomic sequences to be identified as probable members of a large and diverse class of cellular phosphoesterases.
supported by grant LM05798 from the National Library of Medicine

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

Back to Publications Page