Comprehensive Superfamily and Function Classification of Protein Sequences

Huang, H., Barker, W.C., Yeh, L.S., and Wu, C.H.
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 not only by domain and motif identification, but also 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 facilitated by the PIR Annotation and Similarity Database, which includes a pre-computed FASTA Database, and the PIR HMM Homology Domain Database. Annotators, with the help of the Web interfaces, assign the superfamily names, verify membership, and identify distantly related members. Currently, we have classified about 70% of the 240,000 PIR-PSD entries into 33,000 superfamilies. While continuing to automatically classify new protein sequences, we are concentrating manual curation on functionally important superfamilies to further develop our system into an effective tool for function classification. Examples of such application on phosphatase superfamilies will be shown.
Supported by NLM grant LM05798 and NSF grant DBI-9974886

Presented at the Fifth Annual Conference On Computational Genomics, Baltimore, Maryland, 2001

Back to Publications Page