Integrated System for Comprehensive Classification of Protein Sequences

H. Huang, W. C. Barker, 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

ABSTRACT

Many genomic sequences have been annotated on the basis of the highest scoring match in a protein sequence database search. This can be misleading when the match is only between portions of the sequences, which may be a common domain. 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 family and superfamily organization of the PIR-International Protein Sequence Database 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 interface for rapid and accurate classification of large numbers of protein sequences into comprehensive and non-overlapping families and superfamilies. Facilitating the classification is the PIR superfamily information database, which includes superfamily summary information, family summary information, sequence member information, domain arrangement and statistics that measure the goodness of the superfamiliy. This system not only allows rapid the classification but also information retrieval and accurate annontation.

(PIR is supported by grant LM05798 from the National Library of Medicine, USA)

Presented at the 13th International Congress for Methods of Protein Structure Analysis (MPSA), 2000.

Back to Publications Page