PIRSF Protein Classification System (PCS): A Protein Classification, Curation And Annotation Tool

Mazumder R, Hua H, Huang H, Natale DA, Nikolskaya A, Wu CH

The PIRSF Protein Classification System (PCS) allows classification of proteins into a family hierarchy using curator-guided clustering and other built-in tools. We will demonstrate how PCS facilitates large-scale functional annotation by classifying proteins belonging to i) a specific metabolic pathway and ii) a specific genome. The PCS demo is intended for biologists and computer scientists who are interested in protein classification.

A general approach for functional characterization of an unknown protein is to infer function based on the annotation of a "best-hit" sequence in a protein database. This method is susceptible to error because the best hit may not have any information or —worse— may be missannotated. This type of error can be avoided by using a manually-curated classification database such as PIRSF (http://pir.georgetown.edu/pirsf). The advantage of using such a curated database is conferred by the ability to use the collected evidence available from multiple related proteins. Specifically, one can evaluate the relationship of a query sequence with the rest of the members of a family in terms of protein name, sequence diversity, domain architecture and taxonomic distribution. Accordingly, it is ideal for large-scale functional annotation, comparative analysis of protein families and other scientific and bioinformatics analysis.

The PCS was developed at Protein Information Resource (PIR) to facilitate manual classification of proteins into families and superfamilies (PIRSFs) based on analysis of sequence, domain architecture, and additional integrated information available from comprehensive cross-references. This system consists of four components—the underlying database, information retrieval system, sequence analysis tools, and graphical interface for data visualization and family editing. PCS is currently being used at PIR to classify UniProtKB (http://www.uniprot.org/) proteins.