PIRSF Protein Family Classification System

Anastasia Nikolskaya, Sehee Chung, Hongzhan Huang, Raja Mazumder, Darren Natale, Lai-Su Yeh, Cathy H. Wu
Protein Information Resource, Georgetown University Medical Center, Washington, DC 20057

The PIRSF (SuperFamily) protein classification system uses a network structure with multiple levels of sequence diversity from superfamilies to subfamilies to reflect the evolutionary relationship of full-length proteins and domains. The primary PIRSF classification unit is the homeomorphic family, whose members are both homologous (evolved from a common ancestor) and homeomorphic (sharing full-length sequence similarity and a common domain architecture).

Protein family classification provides effective means for large-scale genome annotation and biological knowledge discovery based on the information embedded within families of homologous sequences and their structures. Protein classification based on full-length proteins, including the preservation of domain architecture, allows for accurate annotation of both generic biochemical and specific biological functions of the protein families. This, in turn, facilitates propagation of annotations to the individual protein members of the family and standardization of protein annotation. The multiple levels of sequence diversity, from superfamilies to subfamilies, reflect different degrees of functional granularity and, thereby, allow more accurate propagation of annotation and the development of standard protein nomenclature and ontology. The PIRSF family classification is central to the functional annotation of proteins in the UniProt Knowledgebase (UniProtKB) for both automatic annotation in the UniProtKB/TrEMBL section and literature-based manual annotation in the UniProtKB/Swiss-Prot section.

The PIRSF system allows associative analysis using information on protein sequence, structure, function, and other systems biology information. This integrative approach has led to novel predictions and functional inference for previously uncharacterized proteins, to detection and correction of genome annotation errors, as well as to enhanced understanding of structure, function, and evolutionary relationships. Functional predictions for the uncharacterized protein families are done based on detailed sequence analysis, domain organization, context analysis, phyletic pattern, and other information. PIRSF families are manually curated for membership, annotation of specific biological functions, biochemical activities, and sequence features. Systematic PIRSF family curation follows four complementary approaches—domain-based, function-based, pathway-based, or lineage-based—depending upon priorities and the special properties of the protein families and subsystems.

The PIRSF system consists of two data sets: non-curated clusters and curated families. Currently, about half of UniProtKB sequences are classified into over 29,000 non-curated clusters, including single-member clusters. The non-curated clusters are computationally defined using both pairwise-based parameters and cluster-based parameters. Systematic family curation is being conducted in a two-tier process to improve the quality of automated classification, with over 4,200 “first-tier” (preliminarily-curated) and 1,700 “second-tier” (fully-curated) families, respectively. The first-tier curation provides membership and domain architecture characteristic of the family, while the second-tier curation provides additional annotation, including family name, parent-child relationship, family description, and bibliography. All UniProtKB proteins will be classified in the PIRSF framework. Homologous protein families are defined systematically in an iterative mode that couples manual analysis with computer-assisted clustering and information retrieval.

The new PIRSF curation platform connects a set of analysis and visualization tools and a DAG editor to maximize throughput and minimize routine, error-prone tasks, thereby allowing scientists to provide richly and accurately curated network of protein families. Integrated tools include PIRClust (iterative BlastClust) for sequence clustering, CLustalW for multiple sequence alignment with neighbor-joining phylogenetic tree, a PIR taxonomy tree browser, and the SEED program for genome context and subsystem analysis.

The curation platform is implemented in an n-tier J2EE application composed of a Java Web Start (heavyweight) and browser (lightweight) client tier, a Struts-based web tier with Model, View, Controller (MVC) design, a Data Access Object (DAO) layer and an Oracle database layer. Extensive object-oriented modeling and relational mapping have been performed to obtain application programming interface (API) which makes the application reusable and extensible. Java Web Start technology powered by Sun Microsystems ensures users to get the most recent application on their desktop in system- and language-independent fashion, eliminating the burden of software installation and update. The PIR tree and alignment viewer and DAG editor are such applications embedded within the system. Dynamic JNLP generator passes application parameters to clients at runtime.

PIR has engaged the research community for collaborative expert curation of protein families to enhance both annotation quality and productivity. The PIRSF family curation interface will be made available to collaborating researchers, who will be able to access the PIRSF classification system from a Web browser at any time, always using the most current versions of software tools and data. Interested collaborators can register (pirmail@georgetown.edu) to access the Web-based family curation system via a sign-on system with role-based authentication process. General users can freely use both lightweight and heavyweight client software to search, retrieve and analyze the classified information. Authorized users have access to additional resources and can create, update and save their work into the Oracle database. Any critical modification on the database is audited, versioned and checked for possible conflicts with other curator's work. Client interface is highly configurable and customizable in such a way that users can adjust fonts and colors, choose table columns of interest, run bi-directional sorts, export to and import from famous data formats.

The PIRSF system is integrated with other family, function, and structural classification schemes, and is accessible at http://pir.georgetown.edu/pirsf/ for report retrieval and sequence classification.