The PIRSF system for protein classification, rule-based annotation, and ontology mapping

Darren A. Natale, Cecilia Arighi, Winona Barker, Zhangzhi Hu, Hongzhan Huang, Robel Kahsay, Raja Mazumder, Anastasia Nikolskaya, Sona Vasudevan, C. R. Vinayaka, Lai-Su Yeh, Cathy H. Wu
(*PIR/GUMC*)

The function of characterized proteins is often inferred based on similarity to annotated proteins in sequence databases. Though powerful, this method is prone to errors that propagate throughout molecular databases, including erroneous annotation, under-identification (failure to provide the most specific information) and over-identification (providing too-specific information). These problems can be addressed by using PIRSF--a curated, hierarchical, whole-protein classification database--especially in conjunction with a rule-based system designed specifically for large-scale annotation of individual proteins. This same classification system can also facilitate connections between the three GO vocabularies or other ontologies.

PIRSF: hierarchical whole-protein classification

Classification. Instead of relying on the (hopefully) accurate annotation of a single (hopefully related) protein (usually, the BLAST best hit), using curated classification databases allows reliance on the collected wisdom of multiple proteins, or at least the assurance that the members are truly related.

Whole proteins. Mostly, whole proteins equal to the sum of their parts. However, this is not always the case. For example, the very reasonable “glycerol-3-phosphate dehydrogenase (anaerobic), subunit C” reduces to “Cysteine-rich iron-sulfur binding protein” when its component domains alone are considered. A multi-domain protein with only one domain described may be under-annotated; conversely, a single-domain protein may hit proteins of much longer length (and likely different function). The use of a whole protein classification database, combined with an insistence that predicted members of a given family exhibit (near) end-to-end similarity, obviates such problems.

Hierarchies. The annotation power of protein classification databases is made more powerful if a single database contains families with progressively greater levels of similarity (that is, hierarchies). Theoretically, one query protein could be confidently predicted to be a member of a parent family, but not a child family, while a different query might be confidently assigned to both levels. Propagating the most-specific possible annotation can prevent over- or under-annotation.

The PIRSF system. The PIRSF protein classification system combines all of the approaches described above, providing protein classification from superfamily to subfamily levels in a network structure based on evolutionary relationships of whole proteins. PIRSF classification, which considers both full-length similarity and domain architecture, discriminates between single- and multi-domain proteins where functional differences are associated with the presence or absence of one or more domains. Furthermore, hierarchical classification allows annotation of both generic biochemical and specific biological functions for uncharacterized sequences.

PIR Rules for automated annotation

The PIRSFs are well suited to large-scale protein annotation, affording more robust propagation of information than a simple best-hit approach. However, it is still possible to further refine the system for large-scale automatic annotation by constructing sets of condition/action (if/then) statements into “annotation rules.” The conditions can range from the sequence-based, such as “member of family X,” or organism-based, such as “member of taxonomic lineage A.” The action would be the propagation of appropriate information to the query protein.

Advantages of rule-based annotation. Annotation rules add significant advantages when used in conjunction with protein classification systems for the automated propagation of information from a family to an individual protein:

- Increased specificity. The division of families into subfamilies based on whole-protein similarity is difficult, if not impossible, for proteins with different substrate specificities when the specificity is encoded in a very small number of residues. However, rules can test for known amino acid combinations that confer particular specificity.
- Maintenance. Maintaining a single rule for multiple proteins is easier than maintaining the individual proteins. The annotation of proteins that fit a particular rule can be periodically updated to reflect changes in the rule actions.
- More annotation fields. Ease of maintenance allows flexibility in the number of fields that can be "touched" by automated means. These include not only protein names, but other important annotation fields, including position-specific sequence features, EC name and number, keywords, references, and GO terms.
- Standardization. The uniform application of a rule to proteins in a given family, by definition, will create uniform annotation and a kind of controlled vocabulary to significantly aid text-based searches.
- Evidence attribution. Rules can themselves be annotated with information that describes the rule source and whether the propagatible information is based on experimental evidence or computational prediction, thus providing an effective means to avoid misinterpretation of annotation information and propagation of annotation errors.
- Validation. Annotation rules can also be used to flag unreliable information through "caution" statements.

Annotation rules at PIR. Annotation can be reliably propagated from sequences containing experimentally determined properties to closely-related homologous sequences based on curated PIRSF families. Two types of PIR rules are manually defined and curated. PIR Site Rules focus on sequence-specific features, such as active sites, binding sites, and modified or other functionally important residues. PIR Name Rules propagate names, synonyms and acronyms, EC number, GO terms, and function, pathway, and caution statements. The Name Rules provide the means to account for taxonomically restricted names (or activities) or functional variations within one PIRSF, including instances where a protein lacks the active site residue(s) necessary for enzymatic activity.

PIRSF complements GO

A PIRSF classification-based protein ontology can complement GO concepts by identifying missing GO branches/nodes and linking GO terms among the three vocabularies (i.e., molecular function, biological process, and cellular component). We found that a majority of curated PIRSF families map to GO leaf nodes, and many also share common GO leaf nodes. The PIRSF associations to GO nodes allow us to examine whether certain GO subtrees might need expansion if GO concepts are too broad and to identify missing GO nodes when entire groups of superfamilies cannot be mapped to existing GO terms. PIRSF classification can also provide links between the three GO vocabularies, each of which presently has its own hierarchical organization with no relationships inter-connecting them.