Large-scale, classification-driven, rule-based functional annotation of proteins

Darren A. Natale, C. R.Vinayaka and Cathy H. Wu
Department of Biochemistry and Molecular Biology, Georgetown University Medical Center, Washington, DC, USA

Keywords: database, protein classification, PIRSF, rule-based annotation, sequence analysis

Abstract

Experimentally-verified information on protein function lags far behind the rapid accumulation of protein sequences. The simple approach to propagating information from characterized proteins to unknown proteins—namely, by sequence similarity search against databases of individual proteins—may fail to produce accurate results, and typically is used to transfer only protein name information. A more accurate, consistent, and comprehensive approach for large-scale automated annotation makes use of protein family classification-driven rules. Unannotated proteins that satisfy a set of conditions for a particular rule can be annotated with the information appropriate for that rule. The approach leads to facile, accurate prediction and functional inference for uncharacterized proteins, allows systematic detection of genome annotation errors, and provides sensible propagation and standardization of protein annotation, including position-specific sequence features, protein names and synonyms, and Gene Ontology terms. Rule-based annotation will be discussed in the context of the PIRSF protein classification system, PIRNR Name Rule system, and the PIRSR Site Rule system.

Introduction

The high-throughput genome projects have resulted in a rapid accumulation of predicted protein sequences for a large number of organisms. Meanwhile, scientists have begun to tackle protein functions and other complex regulatory processes using global-scale data generated at various levels of biological organization, ranging from genomes and proteomes to metabolomes (metabolites synthesized by a biological system) and physiomes (the physiological dynamics of whole organisms). To fully realize the value of the data, scientists need to understand how proteins function in making up a living cell. With experimentally-verified information on protein function lagging behind, bioinformatics methods are needed for reliable and large-scale functional annotation of proteins.

A general approach for functional characterization of unknown proteins is to infer function based on sequence similarity to annotated proteins in sequence databases. While this is a powerful method, numerous genome annotation errors have been detected, many of which have been propagated throughout molecular databases. There are several sources of errors (Galperin and Koonin, 1998; Koonin and Galperin, 2003). Errors often occur when identification is made based on local domain similarity or similarity involving only parts of the query and target molecules. Furthermore, the similarity may be to a known domain that is tangential to the main function of the protein or to a region with compositional similarity, such as transmembrane domains. The best hit may be statistically insignificant. Errors also occur when the best hit entry is an uncharacterized or poorly annotated protein, or is itself incorrectly predicted, or simply has a different function. Aside from erroneous annotation, database entries may be under-identified, such as a "hypothetical protein" with a convincing similarity to a protein or domain of known function, or may be over-identified, such as ascribing a specific enzyme activity when a less specific one would be more appropriate.

These problems can be addressed by using a curated, hierarchical, whole-protein classification database, especially in conjunction with a rule-based system designed specifically for large-scale annotation. Annotation rules define not only the conditions that must be met, but also the fields (including function, ontology, and site-specific features) that could be confidently propagated.

Hierarchical whole-protein classification

The advantage of protein classification. There is an immediate advantage to using a protein classification database as a basis for annotation. This advantage is conferred by "strength in numbers." 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. (For the sake of this discussion, it is assumed that the query protein has been accurately assigned to a protein family of interest. This, typically, is done via specialized BLAST search, (e.g., the COG database, Tatusov *et al*., 2003), HMMer search (e.g., Pfam, Bateman *et al*., 2004), or combination of both (e.g., PIRSF, Wu *et al*., 2004a)). *Streptococcus agalactiae* protein gbs1797 (UniProt accession Q8E3G2) provides a good example of the usefulness of protein families for annotation. It is likely that the simple "Hypothetical protein" annotation attached to this protein derived from the fact that its best hit has the same annotation. However, long before this protein was submitted to the sequence databases, the COG database had its predicted family annotated as "Galactose-1-phosphate uridylyltransferase" (COG1085). Similarly, *Mannheimia succiniciproducens* "Hypothetical protein MS0372" (UniProt accession Q65VN1) is predicted to be a member of PIRSF family PIRSF000378 ("Glycyl radical cofactor protein YfiD"), PFAM family PF01228 ("Glycine radical"), and InterPro family IPR001150 ("Formate C-acetyltransferase glycine radical"), all of which were available at the time of submission. Thus, several of the problems associated with propagating meaningful annotation to new proteins, including mis-annotated or statistically insignificant best hits, are averted by the use of protein classification databases.

Whole proteins. Are whole proteins equal to the sum of their parts? Mostly, yes. However, this is not always the case, at least for annotation purposes. For example, PIRSF000142 contains "glycerol-3-phosphate dehydrogenase (anaerobic), subunit C" proteins. The annotation of such proteins, if one were to rely on predominantly domain-centric databases alone, would likely be "Cysteine-rich iron-sulfur binding protein" (based on hits to PF02754, "Cysteine-rich domain" and PF00037, "4Fe-4S binding domain"). Worse yet are the possibilities that a protein is composed of multiple domains, but only one is described (thereby causing an under-annotation by means of omission), or the converse, where a protein is composed of only one domain but hits proteins of much longer length and likely different function (see Barker *et al*., 2005, for examples). The use of a whole protein classification database, such as PIRSF, combined with an insistence that predicted members of a given family exhibit (near) end-to-end similarity, obviates such problems. Other predominantly whole protein classification databases (which may or may not also contain domain families) include COGs, TIGRFAMs (Haft *et al*., 2003), PANTHER (Mi *et al*., 2005), and HAMAP (Gattiker *et al*., 2003).

Hierarchies. The annotation power of protein classification databases is rendered even more powerful if a single database contains families with progressively greater levels of similarity (that is, hierarchies), or if different databases (with different levels) are searched. Such databases, most notably PIRSF and the database-integrating InterPro (Mulder *et al*., 2005), allow annotation at an appropriate level of specificity. 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. In such cases, the most-specific possible annotation could be propagated. For example, PIRSF001370, a large family that contains acetolactate synthase-like thiamine diphosphate-dependent enzymes, can be further divided into seven subfamilies. Two *Pseudomonas putida* members of this family (UniProt accessions Q88DY8 and Q88N22) are annotated as the large subunit of acetolactate synthase (the latter as putative). However, only Q88DY8 belongs to subfamily PIRSF500108, which contains verified acetolactate synthase large subunits. Q88N22 can only be confidently assigned to the parent family. Pending confirmation of activity, "thiamine diphosphate-dependent enzyme" would be a safer annotation. As these examples demonstrate, the use of a hierarchical database helps prevent over- or under-annotation (see Barker *et al*., 2005, for further examples).

The PIRSF system. The PIRSF protein classification system combines all of the approaches described above (Wu *et al*., 2004a). The system provides protein classification from superfamily to subfamily levels in a network structure based on evolutionary relationships of whole proteins. Such classification allows identification of probable function for uncharacterized sequences. 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, classification based on whole proteins, rather than on the component domains, allows annotation of both generic biochemical and specific biological functions.

In this system, unassigned proteins are tested against a set of hidden Markov models (HMMs) representing curated full-length protein families at different hierarchical levels. The following criteria, used by PIRSF Scan (http://pir.georgetown.edu/pirsf) , must be met:

1) The query protein must be recognized by an HMM with a score that falls above the minimum threshold score appropriate for that family. The threshold is based on the minimum score exhibited by current curated members of the family.
2) The length of the query protein must fall within the range exhibited by true members of the family. Both the score and length thresholds are adjusted based on characteristics of the protein family—that is, by taking into account standard deviations—to enable the scan program to capture divergent members for further testing.
3) The set of BLAST hits for the query must be assigned predominantly to the same PIRSF as predicted by HMM. Here, the "majority rule" principle is used, meaning that the query must hit at least 10 or one-third of the members of the PIRSF.

Fine tuning of the match parameters is an ongoing process. Exact parameters used are described in the help document linked at the URL given above.

Rule-based automated annotation

Curated, hierarchical, whole-protein classification databases are, by design, well suited to large-scale protein annotation. First, the classification of whole proteins—not parts—enables specific biological function to be accurately propagated rather than only generic biochemical function (or worse, inaccurate biological function). Second, a hierarchical system of classification allows the propagation of generic biochemical function when a protein cannot confidently be assigned to a sub-family with known specific biological function. However, it is still possible to further refine large-scale automatic annotation systems. Such refinement is afforded by using annotation rules.

Annotation rules defined. An annotation rule is a set of condition/action statements. The conditions can range from the sequence-based, such as "member of family X," "contains domain Y," or "motif Z present," to the organism-based, such as "member of taxonomic lineage A" or "encodes metabolic function B." The action, typically, is the propagation of appropriate information to the query protein, but it can also include the flagging of pre-existing annotation as inappropriate. Rules are most useful (and accurate) when associated with a particular family or set of families. For example, a rule written to propagate the name "Pyruvate (flavodoxin) dehydrogenase" to proteins that match its active site motif could end up wrongly propagating that term to Midasin or other inappropriate proteins. However, if this same rule were applied only to members of the appropriate family (e.g, PIRSF000159), then only suitable proteins would be so annotated, assuming they match the motif.

Annotation rules have thus far been developed by the three members of the UniProt Consortium—the Protein Information Resource (PIR), the Swiss Institute of Bioinformatics (SIB), and the European Bioinformatics Institute (EBI). PIR has developed manually-curated Site Rules (PIRSRs) and Name Rules (PIRNRs), each of which (described in detail below) is based on curated protein families of the PIRSF system. SIB has developed curated rules based on the curated HAMAP family system (Gattiker *et al*., 2003). EBI has multiple systems—the manually curated RuleBase system (Biswas *et al*., 2002) and the automatically generated Spearmint (Kretschmann *et al*., 2001) and Xanthippe (Wieser *et al*., 2004) rules. Each of these rule sets are integrated into the annotation pipeline for proteins in the UniProt Knowledgebase (Bairoch *et al*., 2005).

Advantages of rule-based propagation of functional 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.** For example, both archaea and eukarya contain homologs to the DNA polymerase sliding clamp (PIRSF002090), commonly called Proliferating Cell Nuclear Antigen (PCNA). However, though the designation PCNA is perfectly reasonable for eukarya, it is not for archaea. Furthermore, the archaeal and the eukaryal versions are not easily separable based on sequence. Therefore, the only recourse for proper naming would be a taxonomy-based rule. Another example is provided by PIRSF000532, a protein family composed of members with the same general function but different specificities. The proteins are phosphofructokinases that are either ATP-dependent, pyrophosphate-dependent, or of unknown dependency. The specificity is encoded in a very small number of residues (Bapteste *et al.*, 2003). Multiple instances of convergent evolution render the division of this family into subfamilies based on whole-protein similarity to be difficult if not impossible. However, rules can test for the known amino acid combinations conferring ATP or pyrophosphate dependence, thereby allowing the propagation of “ATP-dependent phosphofructokinase” or “Pyrophosphate-dependent phosphofructokinase” as appropriate. In addition, a fallback rule can be formulated such that entries failing both of the specific rules could be named the less-specific, but still accurate, “Phosphofructokinase.”

- **Maintenance.** This attribute works on two levels. First, maintaining a single rule for multiple proteins is easier than maintaining those proteins individually. Second, at least as applied in UniProt, the annotation of proteins that fit a particular rule can be periodically updated to reflect changes in the rule actions. The reapplication of rule actions means that entries previously annotated based on a particular rule would immediately be fitted with the new information from the updated rule.
- **More annotation fields.** The ease of maintainance fosters an increase in the number of functional annotation fields that can be “touched” by automated means. Most current methods for applying automatic annotation to individual proteins focuses specifically on the protein name (even though, in principle, other fields can be annotated as well; in practice, this is not done). However, rules afford an easy mechanism for more accurate propagation of other important annotation fields, such as position-specific sequence features, Enzyme Commission (EC) name and number, keywords, references, and Gene Ontology (GO) terms. Function-based classifications, such as EC and GO, provide useful complements to sequence-based classifications enabling, for example, studies on analogous enzymes (Galperin *et al.*, 1998).
- **Standardization.** The uniform application of a rule to proteins in a given family, by definition, will create a uniformity in annotation (when warranted). When applied to protein names, such annotation becomes another controlled vocabulary, thereby significantly aiding text-based searches.
- **Evidence attribution.** Rules can themselves be annotated with information that describes the source for the rule and whether the propagatible information is based on experimental evidence or computational prediction. Storing rules with unique identifiers allows information propagated by a rule to be tagged with the rule identifier. This, in effect, forms the basis for evidence attribution describing data source, types of evidence, and methods for annotation. Such evidence attribution provides an effective means to avoid misinterpretation of annotation information and propagation of annotation errors.
- **Validation.** A large amount of inaccurate information can be (and has been) propagated to proteins using non-stringent criteria. These erroneous data are difficult to find, so nearly impossible to eradicate completely. However, annotation rules—used to propagate reliable information—can also be used to flag unreliable information through “caution” statements. Each of the annotation rule systems indicated above has this component; however, the Xanthippe system performs this function exclusively.

Annotation rules at PIR. The PIRSF classification serves as the basis for a rule-based approach that provides standardized and rich automatic functional annotation. In particular, annotation can be reliably propagated from sequences containing experimentally determined properties to closely related homologous sequences based on curated PIRSF families. PIR rules are manually defined and curated for several annotation fields, as described below. Two types of annotation rules have been developed. PIR Site Rules focus on sequence-specific features (UniProt "FT" lines). Originally designed to capture and propagate protein name (UniProt "DE" line) information, PIR Name Rules have expanded in scope to include synonyms and acronyms, Enzyme Commission (EC) name and number, Gene Ontology terms, and comment fields such as Function, Pathway, and Caution.

PIRSR Site Rules. To assure correct functional assignments, protein identifications must be based on both global (whole protein) and local (domain and motif) sequence similarities. Position-specific Site Rules enable the annotation of active site residues, binding site residues, modified residues, or other functionally important amino acid residues. To exploit known structure information, Site Rules are defined starting with PIRSF families that contain at least one known 3D structure with experimentally-verified site information. The active site information on proteins is taken from the PDB (Bourne *et al*., 2004) SITE records, the LIGPLOT of interactions available in PDBSum database (Laskowski, 2001), and published scientific literature. The dataset of catalytic residues (Bartlett *et al*., 2002) is used as an authoritative source for catalytic residues in enzyme active sites.

Site rule curation involves manually editing multiple sequence alignments of representative protein family members (inlcuding the template PDB entry), building hidden Markov models (HMMs) from the conserved regions containing the functional site residues, and visualizing sequences and structures. The rules are defined using appropriate syntax and controlled vocabulary for site description and evidence attribution. The profile HMM thus built allows one to map functionally important residues from the template structure to other members of the PIRSF family that do not have a solved structure. To avoid false positives, site features are only propagated automatically if all site residues match perfectly in the conserved region by aligning both the template and target sequences to the profile HMM using HmmAlign (Eddy *et al*., 1995). Potential functional sites missing one or more residues or containing conservative substitions are only annotated after expert review with evidence attribution. Table 1 shows six example Site Rules, three for PIRSF000077 and three for PIRSF000097.

"Table 1 near here"

Automatic annotation generated by the PIR Site Rules will be attributed with the rule ID, which links to the rule report containing additional information. For Site Rules, the rule report will include multiple sequence alignments with highlighted site residues (Figure 1).

"Figure 1 near here"

PIRNR Name Rules. To capture the full annotation capability of PIRSFs, PIR Name Rules (PIRNRs) were developed. As with Site Rules, each Name Rule is defined with conditions for propagation (Table 1). While most Name Rules assign protein names to all family members, many families require more specialized rules with additional conditions to propagate appropriate names and avoid over-identification or under-identification. In addition to the PCNA example of taxonomically restricted names (or activities), PIR Name Rules provide the means to account for functional variations within one PIRSF, including instances where a protein lacks the active site residue(s) necessary for enzymatic activity, and cases where evolutionarily-related proteins have known differences in biochemical activities (such as the phosphofructokinases of PIRSF000532) or domain organization.

PIRNRs are part of a two-tier system. The first tier is the "zero-level" rule. Such rules are designed without the ability for discrimination. That is, the information propagated by such a rule will apply to all members of the PIRSF to which the rule belongs (each PIRNR will apply only to members of its cognate PIRSF, and to no other PIRSF members). This ensures that every classified protein will get some annotation (assuming any information is known for the PIRSF members, and such information is deemed applicable to all members). The "higher-order" rules will fit only those members that pass additional tests. The additional tests may include taxonomic placement, presence or absence of critical domains, fit to a PIR Site Rule, or presence on manually curated inclusion (false negative) or exclusion (false positive) lists. By definition, higher-order rules are more specific than zero-level rules. Therefore, the propagation system will be designed to prevent propagation based on the zero rule when a higher-order rule applies.

Creating a PIRNR, which seeks to capture known or predicted information about a PIRSF, follows the completed curation of each PIRSF:

1. *Compile known (literature) or predicted (sequence analysis) information about members of the PIRSF of interest.*
2. *Indicate match conditions (higher-order rules only), including*
    a. Taxonomic range to which the rule should apply.
    b. Essential domains (in the form of matches to PFAM).
    c. Essential residues (in the form of matches to PIR Site Rules).
    d. False negatives (list known entries that should get the indicated annotation, but would otherwise fail the rule tests).
3. *Indicate exclusion conditions (higher-order rules only), including*
    a. Taxonomic range that should NOT get propagation.
    b. Domains that should NOT be present.
    c. Residues that should NOT be present.
    d. False positives (list known entries that pass all the rule tests, but should NOT get propagated information).
4. *Indicate propagatible information, including*
    a. Protein name (DE line) and acronym that adheres to UniProt standards.
    b. Alternative names or synonyms used in the literature, with acronyms.
    c. Comments on function, pathway, or cautions about nomenclature (as in the PIRNR025624-1 example; Table 1).
    d. EC number.
    e. GO terms. Only the leaf-most term is propagated.
5. *Outline the scope of the rule, or provide reasons why a particular protein name was chosen as standard for the group.*

Conclusions

Critical to our understanding of biology is accurate and up-to-date information. The process of evolution affords us the ability to make inferences about the nature of the proteins that govern biological processes, since like proteins often perform like (if not exact) functions. Unfortunately, this same process has been far from a smooth transition from state to state. The result is that inferences made about one protein based on similarity to another protein using automated methods are often suspect. This is more than a mere annoyance. The lack of rigorous methods for propagating appropriate information hampers knowledge discovery by either reducing the associations that can be made, or by producing associations that should not be made. However, the recent development of methods for better annotation hold much promise for preventing—and even reversing—the previous trend toward rampant misinformation. The combination of hierarchical, whole-protein classifications and rule-based large-scale annotation pipelines is a significant step in the right direction.

Acknowledgments

The project is supported by grant U01-HG02712 from National Institutes of Health and grant DBI-0138188 from National Science Foundation.

Related articles

g306302

References

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N and Yeh LS (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Research*, **33**, D154-D159.

Bapteste E, Moreira D and Philippe H (2003) Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase. *Gene*, **318**, 185-191.

Barker WC, Mazumder R, Nikolskaya AN, and Wu CH (2005) The PIR SuperFamily (PIRSF) classification system. *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. Proteomics Volume, Subramaniam, S. (Ed.) John Wiley & Sons, Ltd. (article g306302 in press).

Bartlett GJ, Porter CT, Borkakoti N and Thornton JM (2002) Analysis of catalytic residues in enzyme active sites. *Journal of Molecular Biology*, **324**, 105-121.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C and Eddy SR (2004) The Pfam protein families database. *Nucleic Acids Research*, **32**, D138-D141.

Biswas M, O'Rourke JF, Camon E, Fraser G, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, Mittard V, Mulder N, Phan I, Servant F and Apweiler R (2002) Applications of InterPro in protein annotation and genome analysis. *Brief Bioinform*., **3**, 285-295.

Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott JC, Townsend-Merino W, Weissig H, Westbrook J and Berman HM (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Research*, **32**, D223-D225.

Eddy SR, Mitchison G and Durbin R (1995) Maximum Discrimination hidden Markov models of sequence consensus. *Journal of Computational Biology*, **2**, 9-23.

Galperin MY and Koonin EV (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biology*, **1**, 55-67.

Galperin MY, Walker DR and Koonin EV (1998) Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, **8**, 779-790.

Gattiker A, Michoud K, Rivoire C, Auchincloss AH, Coudert E, Lima T, Kersey P, Pagni M, Sigrist CJ, Lachaize C, Veuthey AL, Gasteiger E, and Bairoch A (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput Biol Chem*., **27**, 49-58.

Haft DH, Selengut JD, and White O (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research*, **31**, 371-373.

Koonin EV and Galperin MY (2003) *Sequence – Evolution – Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers: Boston, MA.

Kretschmann E, Fleischmann W and Apweiler R (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920-926.

Laskowski RA (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, **29**, 221-222.

Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, and Thomas PD (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, **33**, D284-D288.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya AN, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJ, Silventoinen V, Studholme DJ, Vaughan R, and Wu CH (2005) InterPro, progress and status in 2005. *Nucleic Acids Research*, **33**, D201-D205.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Wieser D, Kretschmann E and Apweiler R (2004) Filtering erroneous protein annotation. *Bioinformatics*, **20 Suppl 1**, I342-I347.

Wu CH, Yeh L-S, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE, Vinayaka CR, Zhang J and Barker WC (2003a) The Protein Information Resource. *Nucleic Acids Research*, **31**, 345-347.

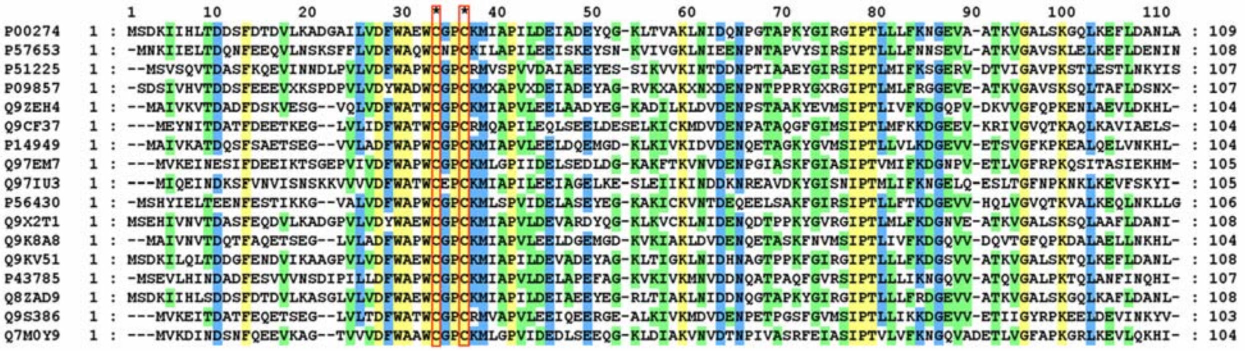
Wu CH, Huang H, Yeh LS and Barker WC (2003b) Protein family classification and functional annotation. *Computational Biology and Chemistry*, **27**, 37-47.

Wu CH, Nikolskaya A, Huang H, Yeh L-S, Natale D, Vinayaka CR, Hu Z, Mazumder R, Kumar S, Kourtesis P, Ledley RS, Suzek BE, Arminski L, Chen Y, Zhang J, Cardenas JL, Chung S, Castro-Alvear J, Dinkov G and Barker WC (2004a) PIRSF family classification system at the Protein Information Resource. *Nucleic Acids Research*, **32**, D112-D114.

Table 1. PIR Site Rules and Name Rules for automated annotation of functional sites and standardized protein names.

Rule Type/ID	Condition	Propagatible Annotation	Evidence
Site rule/ PIRSR000077-1	PIRSF000077 member and HMM site match	Feature: Active site Residues: Cys 33, Cys 36	Template: UniProt:P00274; PDB:2TRX Status: Validated [PMID:2181145]
Site rule/ PIRSR000077-2	PIRSF000077 member and HMM site match	Feature: Site Residues: Gly 34, Pro 35	Template: UniProt:P00274; PDB:2TRX Status: Validated [PMID:9099998]
Site rule/ PIRSR000077-3	PIRSF000077 member and HMM site match	Feature: Site Residue: Asp 27	Template: UniProt:P00274; PDB:2TRX Status: Validated [PMID:9374473]
Site rule/ PIRSR000097-1	PIRSF000097 member and HMM site match	Feature: Active site Residue: Tyr 49	Template: UniProt:P15121; PDB:1US0 Status: Validated [PMID:8245005]
Site rule/ PIRSR000097-2	PIRSF000097 member and HMM site match	Feature: Binding site Residue: His 111	Template: UniProt:P15121; PDB:1US0 Status: Validated [PMID:8245005]
Site rule/ PIRSR000097-3	PIRSF000097 member and HMM site match	Feature: Site Residue: Lys 78	Template: UniProt:P15121; PDB:1US0 Status: Validated [PMID:15146478]
Name Rule/ PIRNR001555-1	PIRSF001555 member	Name: aspartate—ammonia ligase Synonym: asparagine synthetase A EC: aspartate—ammonia ligase (EC 6.3.1.1)	Template: UniProt:P00963 Status: Validated [PMID:1369484]
Name Rule/ PIRNR000881-1	PIRSF000881 member and vertebrates	Name: S-acyl fatty acid synthase thioesterase EC: oleoyl-[acyl-carrier-protein] hydrolase (EC 3.1.2.14)	Template: UniProt:P00633 Status: Validated [PMID:2415525]
Name Rule/ PIRNR000881-2	PIRSF000881 member and not vertebrates	Name: Type II thioesterase EC: thiolester hydrolases (EC 3.1.2.-)	Template: UniProt:Q08788 Status: Validated [PMID:9560421]
Name Rule/ PIRNR025624-1	PIRSF025624 member	Name: ACT domain protein Misnomer: chorismate mutase	Status: Predicted

Figure 1. Multiple sequence alignment of PIRSF family Thioredoxin (PIRSF000077), the two active site cysteines (Site Rule PIRSR000077-1) are boxed in red. Thioredoxins are small redox proteins that catalyze dithiol-disulfide exchange reactions. The source organisms, with the UniProt accession numbers in parenthesis, are *Escherichia coli* (P00274), *Buchnera aphidicola* (P57653), *Porphyra purpurea* (P51225), *Chromatium vinosum* (P09857), *Staphylococcus aureus* (Q9ZEH4), *Lactococcus lactis* (Q9CF37), *Bacillus subtilis* (P14949), *Clostridium acetobutylicum* (Q97EM7), *Clostridium acetobutylicum* (Q97IU3), *Helicobacter pylori* (P56430), *Pseudomonas aeruginosa* (Q9X2T1), *Bacillus halodurans* (Q9K8A8), *Vibrio cholerae* (Q9KV51), *Haemophilus influenzae* (P43785), *Yersinia pestis* (Q8ZAD9), *Listeria monocytogenes* (Q9S386) and *Clostridium pasteurianum* (Q7M0Y9).



```
                 1       10        20        30 *  * 40        50        60        70        80        90        100       110
P00274    1 : MSDKIIHLTDDSFDTDVLKADGAILVDFWAEWCGPCKMIAPILDEIADEYQG-KLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVA-ATKVGALSKGQLKEFLDANLA : 109
P57653    1 : -MNKIIELTDQNFEEQVLNSKSFFLVDFWAQWCNPCKILAPILEEISKEYSN-KVIVGKLNIEENPNTAPVYSIRSIPTLLLFNNSEVL-ATKVGAVSKLELKEFLDENIN : 108
P51225    1 : --MSVSQVTDASFKQEVINNDLPVLVDFWAPWCGPCRMVSPVVDAIAEEYES-SIKVVKINTDDNPTIAAEYGIRSIPTLMIFKSGERV-DTVIGAVPKSTLESTLNKYIS : 107
P09857    1 : -SDSIVHVTDDSFEEEVXKSPDPVLVDYWADWCGPCKMXAPVXDEIADEYAG-RVKXAKXNXDENPNTPPRYGXRGIPTLMLFRGGEVE-ATKVGAVSKSQLTAFLDSNX- : 107
Q9ZEH4    1 : --MAIVKVTDADFDSKVESG--VQLVDFWATWCGPCKMIAPVLEELAADYEG-KADILKLDVDENPSTAAKYEVMSIPTLIVFKDGQPV-DKVVGFQPKENLAEVLDKHL- : 104
Q9CF37    1 : ---MEYNITDATFDEETKEG--LVLIDFWATWCGPCRMQAPILEQLSEELDESELKICKMDVDENPATAQGFGIMSIPTLMFKKDGEEV-KRIVGVQTKAQLKAVIAELS- : 104
P14949    1 : --MAIVKATDQSFSAETSEG--VVLADFWAPWCGPCKMIAPVLEELDQEMGD-KLKIVKIDVDENQETAGKYGVMSIPTLLVLKDGEVV-ETSVGFKPKEALQELVNKHL- : 104
Q97EM7    1 : ---MVKEINESIFDEEIKTSGEPVIVDFWAPWCGPCKMLGPIIDELSEDLDG-KAKFTKVNVDENPGIASKFGIASIPTVMIFKDGNPV-ETLVGFRPKQSITASIEKHM- : 105
Q97IU3    1 : ---MIQEINDKSFVNVISNSKKVVVVDFWATWCEPCKMIAPILEEIAGELKE-SLEIIKINDDKNREAVDKYGISNIPTMLIFKNGELQ-ESLTGFNPKNKLKEVFSKYI- : 105
P56430    1 : -MSHYIELTEENFESTIKKG--VALVDFWAPWCGPCKMLSPVIDELASEYEG-KAKICKVNTDEQEELSAKFGIRSIPTLLFTKDGEVV-HQLVGVQTKVALKEQLNKLLG : 106
Q9X2T1    1 : MSEHIVNVTDASFEQDVLKADGPVLVDYWAEWCGPCKMIAPVLDEVARDYQG-KLKVCKLNIDENQDTPPKYGVRGIPTLMLFKDGNVE-ATKVGALSKSQLAAFLDANI- : 108
Q9K8A8    1 : --MAIVNVTDQTFAQETSEG--LVLADFWAPWCGPCKMIAPVLEELDGEMGD-KVKIAKLDVDENQETASKFNVMSIPTLIVFKDGQVV-DQVTGFQPKDALAELLNKHL- : 104
Q9KV51    1 : MSDKILQLTDDGFENDVIKAAGPVLVDFWAEWCGPCKMIAPILDEVADEYAG-KLTIGKLNIDHNAGTPPKFGIRGIPTLLLFKDGSVV-ATKVGALSKTQLKEFLDANL- : 108
P43785    1 : -MSEVLHINDADFESVVVNSDIPILLDFWAPWCGPCKMIAPVLDELAPEFAG-KVKIVKMNVDDNQATPAQFGVRSIPTLLLIKNGQVV-ATQVGALPKTQLANFINQHI- : 107
Q8ZAD9    1 : MSDKIIHLSDDSFDTDVLKASGLVLVDFWAEWCGPCKMIAPILDEIAEEYEG-RLTIAKLNIDDNQGTAPKYGIRGIPTLLLFRDGEVV-ATKVGALSKGQLKAFLDANL- : 108
Q9S386    1 : ---MVKEITDATFEQETSEG--LVLTDFWATWCGPCRMVAPVLEEIQEERGE-ALKIVKMDVDENPETPGSFGVMSIPTLLIKKDGEVV-ETIIGYRPKEELDEVINKYV- : 103
Q7M0Y9    1 : ---MVKDINDSNFQEEVKAG--TVVVDFWAAWCGPCKMLGPVIDEDLSEEQG-KLDIAKVNVDTNPIVASRFEIASIPTVLVFKNGQVADETLVGFAPKGRLKEVLQKHI- : 104
```