Available online at www.sciencedirect.com

SCIENCE DIRECT®

Computational Biology and Chemistry

ELSEVIER

Computational Biology and Chemistry 27 (2003) 37–47

www.elsevier.com/locate/compbiolchem

Protein family classification and functional annotation

Cathy H. Wu *, Hongzhan Huang, Lai-Su L. Yeh, Winona C. Barker

Georgetown University Medical Center and National Biomedical Research Foundation, 3900 Reservoir Road, NW, Box 571455, Washington, DC 20057-1455, USA

Received 31 October 2002; received in revised form 13 November 2002; accepted 14 November 2002

Abstract

With the accelerated accumulation of genomic sequence data, there is a pressing need to develop computational methods and advanced bioinformatics infrastructure for reliable and large-scale protein annotation and biological knowledge discovery. The Protein Information Resource (PIR) provides an integrated public resource of protein informatics to support genomic and proteomic research. PIR produces the Protein Sequence Database of functionally annotated protein sequences. The annotation problems are addressed by a classification-driven and rule-based method with evidence attribution, coupled with an integrated knowledge base system being developed. The approach allows sensitive identification, consistent and rich annotation, and systematic detection of annotation errors, as well as distinction of experimentally verified and computationally predicted features. The knowledge base consists of two new databases, sequence analysis tools, and graphical interfaces. PIR-NREF, a non-redundant reference database, provides a timely and comprehensive collection of all protein sequences, totaling more than 1,000,000 entries. iProClass, an integrated database of protein family, function, and structure information, provides extensive value-added features for about 830,000 proteins with rich links to over 50 molecular databases. This paper describes our approach to protein functional annotation with case studies and examines common identification errors. It also illustrates that data integration in PIR supports exploration of protein relationships and may reveal protein functional associations beyond sequence homology.

Keywords: Data integration; Knowledge base; Protein annotation; Protein classification; Protein database; Protein family

1. Introduction

The high-throughput genome projects have resulted in a rapid accumulation of genome sequences for a large number of organisms. To fully realize the value of the data, scientists need to identify proteins encoded by these genomes and understand how these proteins function in making up a living cell. With experimentally verified information on protein function lagging far behind, computational methods are needed for reliable and large-scale functional annotation of proteins.

A general approach for functional characterization of unknown proteins is to infer protein functions based on sequence similarity to annotated proteins in sequence databases. This complex and ambiguous process is inevitably error prone (Bork and Koonin, 1998). Indeed, numerous genome annotation errors have been detected (Brenner, 1999; Devos and Valencia, 2001), many of which have been propagated throughout other molecular databases. There are several sources of errors. Since many proteins are multifunctional, the assignment of a single function, which is still common in genome projects, results in incomplete or incorrect information. Errors also often occur when the best hit in pairwise sequence similarity searches is an uncharacterized or poorly annotated protein, or is itself incorrectly predicted, or simply has a different function. While assignment of function by sequence similarity is a powerful approach that has led to many scientific discoveries, to avoid errors it must be applied carefully, using a variety of algorithms and databases, coupled with manual curation.

The Protein Information Resource (PIR) (Wu et al., in press) provides an integrated public resource of protein informatics to support genomic and proteomic research and scientific discovery. PIR produces the

* Corresponding author. Tel.: +1-202-687-1039; fax: +1-202-687-1662.
E-mail address: wuc@georgetown.edu (C.H. Wu).

1476-9271/02/$ - see front matter © 2002 Elsevier Science Ltd. All rights reserved.
PII: S1476-9271(02)00098-1

Protein Sequence Database (PSD) of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* edited by Dayhoff (1965–1978). The annotation problems are addressed by a classification-driven and rule-based method with evidence attribution, coupled with an integrated knowledge base system being developed. The knowledge base consists of two new databases to provide a comprehensive protein sequence collection and extensive value-added protein information, as well as sequence analysis tools and graphical interfaces. This paper describes our approach to the functional annotation of proteins with case studies and illustrates how data integration in PIR supports exploration of protein functional associations.

2. Classification-driven and rule-based annotation with evidence attribution

2.1. Protein family classification

Classification of proteins provides valuable clues to structure, activity, and metabolic role. Protein family classification has several advantages as a basic approach for large-scale genomic annotation: (1) it improves the identification of proteins that are difficult to characterize based on pairwise alignments; (2) it assists database maintenance by promoting family-based propagation of annotation and making annotation errors apparent; (3) it provides an effective means to retrieve relevant biological information from vast amounts of data; and (4) it reflects the underlying gene families, the analysis of which is essential for comparative genomics and phylogenetics.

In recent years, a number of different classification systems have been developed to organize proteins. Scientists recognize the value of these independent approaches, some highly automated and others curated. Among the variety of classification schemes are: (1) hierarchical families of proteins, such as the superfamilies/families (Barker et al., 1996) in the PIR-PSD, and protein groups in ProtoMap (Yona et al., 2000); (2) families of protein domains, such as those in Pfam (Bateman et al., 2002) and ProDom (Corpet et al., 2000); (3) sequence motifs or conserved regions, such as in PROSITE (Falquet et al., 2002) and PRINTS (Attwood et al., 2002); (4) structural classes, such as in SCOP (Lo Conte et al., 2002) and CATH (Pearl et al., 2001); as well as (5) integrations of various family classifications, such as iProClass (Huang et al., in press) and InterPro (Apweiler et al., 2001). While each of these databases is useful for particular needs, no classification scheme is by itself adequate for addressing all genomic annotation needs.

The PIR superfamily/family concept (Dayhoff, 1976), the original such classification based on sequence similarity, is unique in providing comprehensive and non-overlapping clustering of protein sequences into a hierarchical order to reflect their evolutionary relationships. Proteins are assigned to the same superfamily/family only if they share end-to-end sequence similarity, including common domain architecture (i.e. the same number, order, and types of domains), and do not differ excessively in overall length (unless they are fragments or result from alternate splicing or initiators). Other major family databases are organized based on similarities of domain or motif regions alone, as in Pfam and PRINTS. There are also databases that consist of mixtures of domain families and families of whole proteins, such as SCOP and TIGRFAMs (Haft et al., 2001). However, in all of these, the protein-to-family relationship is not necessarily one-to-one, as in PIR superfamily/family, but can also be one-to-many. The PIR superfamily classification is the only one that explicitly includes this aspect, which can serve to discriminate between multidomain proteins where functional differences are associated with presence or absence of one or more domains.

Family and superfamily classification frequently allow identification or probable function assignment for uncharacterized (‘hypothetical’) sequences. To assure correct functional assignments, protein identifications must be based on both global (whole protein, e.g. PIR superfamily) and local (domain and motif) sequence similarities, as illustrated in the case studies below.

2.2. Rule-based annotation

Family and superfamily classification also serves as the basis for rule-based procedures that provide rich automatic functional annotation among homologous sequences and perform integrity checks. Combining the classification information and sequence patterns or profiles, numerous rules have been defined to predict position-specific sequence features such as active sites, binding sites, modification sites, and sequence motifs. For example, when a new sequence is classified into a superfamily containing a ‘ferredoxin [2Fe–2S] homology domain,’ that sequence is automatically searched for the pattern for the 2Fe–2S cluster and if the pattern is found, the feature ‘Binding site: 2Fe–2S cluster (Cys) (covalent)’ is added. Such sequence features are most accurately predicted if based on patterns or profiles derived from sequences closely related to those that are experimentally verified. For example, within the cytochrome *c* domain (PF00034), the ‘CXXCH’ pattern, containing three annotatable residues, is easily identified and the ligands (heme and heme iron) are invariant; however, there is no single pattern derivable for identifying the Met that is the second axial ligand of the heme iron. In contrast, within the many superfamilies containing the calcineurin-like phosphoesterase domain (PF00149), the metal chelating residues, the identity of the bound metal ion, and the catalytic activity are variable. In such a case, automated annotation must be superfamily-specific in order to be accurate. Integrity checks are based on PIR controlled vocabulary, standard nomenclature, and other ontologies. For example, the IUBMB Enzyme Nomenclature is used to detect obsolete EC numbers, misspelled enzyme names, or inconsistent EC number and enzyme name.

Table 1 illustrates how an integrated set of rules can be triggered by family classification to produce dynamic annotation of protein entries. After classification, all SF000460 superfamily members containing positive identifications of Pfam domain PF00343 and PROSITE motif PS00102 are automatically tested (action 2) for the superfamily-tailored pyridoxal-phosphate binding site 'EASG[QT][GS]NM^KXXXN[GR]' (where K is the residue covalently binding the cofactor) and instructions are generated to add the appropriate feature. If all essential sequence and site features are present, the entry title is checked for the string 'phosphorylase (EC 2.4.1.1)' or, even more specifically, 'starch phosphorylase (EC 2.4.1.1) if the organism is a plant (action 4). Then, enzyme-associated keywords such as 'glycosyltransferase' and 'hexosyltransferase' are added (action 5). For superfamily members from animals, we test for the phosphorylase kinase phosphorylation site and add the appropriate feature (action 3). The two new features trigger the addition of keywords 'pyridoxal phosphate' (action 6), and 'phosphoprotein' and 'allosteric regulation' (action 7), respectively.

We derive family-specific patterns for such features from alignments of closely related sequences for which some of the sequences have experimentally determined properties. The rule may further specify other topological constraints for the pattern, such as restricting the annotation of the P-loop feature to the ABC transporter domain regions for the excinuclease ABC chain A superfamily. We look for expected active site and binding site sequence motifs and predict disulfide bonds only by homology within the family or superfamily. As a consequence, we do not annotate signal sequences for nuclear proteins, myristylation sites internally in sequences, phosphorylation sites when there is no evidence that the protein is phosphorylated, carbohydrate-binding sites in cytosolic proteins, etc. Sometimes the concatenation of predicted features in a sequence is so plausible as to justify a functional classification and feature annotation even if there is no family or superfamily member with validated function. For example, a eukaryotic protein containing a predicted signal sequence, followed by several predicted immunoglobulin-like domains, followed by a predicted transmembrane domain, followed by a predicted protein kinase or protein phosphatase domain is very likely a receptor involved in a signal transduction pathway.

Table 1
Classification-driven and rule-based approach for automated and quality annotation

Action	Process	Rule	Description [a]
1	Protein classification	Superfamily	Superfamily: SF000460, phosphorylase Domain: PF00343, carbohydrate phosphorylase Motif: PS00102, phosphorylase pyridoxal-phosphate attachment site

2	Site identification	Feature rule 1	IF pattern: EASG[QT][GS]NM^KXXXN[GR]
			THEN add feature: 'binding site: pyridoxal phosphate (Lys) (covalent)'

3	Site identification	Feature rule 2	IF superfamily member + animal (Metazoa)
			And IF pattern: [KR][KR][KR]QI^S[VIL]RG
			THEN add feature: 'binding site: phosphate (Ser) (covalent) (by phosphorylase kinase) (in phosphorylase *a*)'

4	Protein name checking	Protein name rule	IF: superfamily member + feature rule 1
			THEN use name: 'phosphorylase (EC 2.4.1.1)'
			IF: superfamily member + feature rule 1 + plant (Viridiplantae)
			THEN use name: 'starch phosphorylase (EC 2.4.1.1)'

5	Keyword checking	Keyword rule	IF protein name includes: EC 2.4.1.1
			THEN add keywords: 'glycosyltransferase', 'phosphorylase', 'hexosyltransferase'

6	Keyword checking	Keyword rule	IF: feature rule 1
			THEN add keyword: 'pyridoxal phosphate'

7	Keyword checking	Keyword rule	IF: feature rule 2
			THEN add keywords: 'phosphoprotein', 'allosteric regulation'

[a] The quoted texts are terms in PIR controlled vocabularies.

2.3. Evidence attribution and bibliography mapping

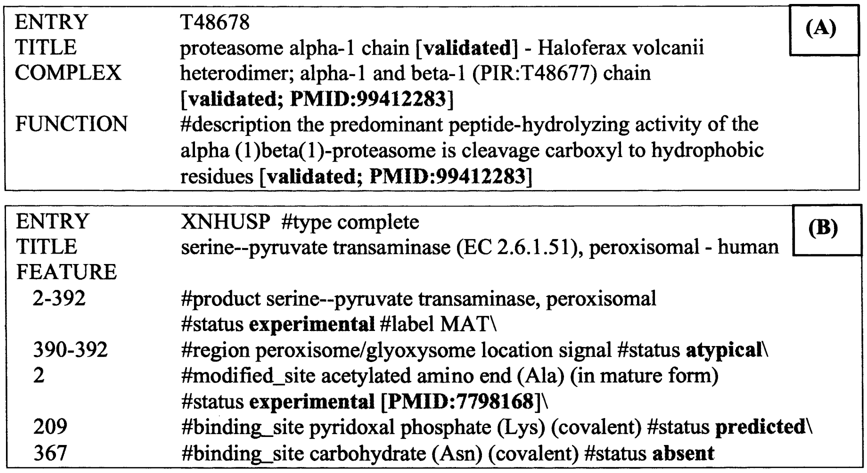
Attribution of protein annotations to validated experimental sources provides effective means to avoid propagation of errors that may have resulted from large-scale genome annotation. To distinguish experimentally verified from computationally predicted data, PIR entries are labeled with status tags of '*validated*', '*similarity*', or '*imported*' in protein title, function, and complex annotations (Fig. 1A). The *validated* function or complex annotation includes hypertext-linked PubMed unique identifiers for the articles in which the experimental determinations are reported. The entries are also tagged with '*experimental*', '*absent*', '*atypical*', or '*predicted*' in feature annotations (Fig. 1B). The first two tags are used to indicate the experimentally determined presence or absence of features. To appropriately attribute bibliographic data to features with experimental evidence, we are conducting a retrospective bibliography mapping. Literature citations within each protein entry are computationally filtered based on both titles and abstracts, using controlled terms describing the experimental features. Subsequently, the filtered papers are manually curated and added to the feature lines as literature attributions.

The amount of experimentally verified annotation available in sequence databases, however, is rather limited due to the laborious nature of knowledge extraction from the literature. Linking protein data to more bibliographic data that describes or characterizes the proteins is crucial for increasing the amount of experimental information and improving the quality of protein annotation. We have developed a bibliography system that provides literature data mining, displays composite bibliographic data compiled from multiple sources, and allows scientists/curators to submit, categorize, and retrieve bibliographic data for protein entries. The submission interface guides users through steps in mapping the citation to protein entries, entering the bibliographic data, and summarizing the contents using categories (such as genetics, tissue/cellular localization, molecular complex or interaction, function, regulation, and disease), with evidence attribution (experimental or predicted) and description of methods. The information page provides literature data mining and displays references collected from curated databases and submitted by users, with PubMed links.

3. Case studies

3.1. IMP dehydrogenase: error propagation to secondary databases

During the PIR superfamily classification and curation process, at least 18 proteins were found to be misannotated as inosine-5′-monophosphate dehydrogenase (IMPDH) or related in various complete genomes. These '*misnomers*,' all of which have been corrected in the PIR-PSD and some corrected in Swiss-Prot/TrEMBL (Bairoch and Apweiler, 2000), still exist in GenPept (annotated GenBank translations) and RefSeq (Pruitt and Maglott, 2001). The misannotation apparently resulted from local sequence similarity to the CBS domain, named for the protein in which it was first described, cystathionine beta synthase, mutations in which cause homocystinuria, an inborn error of metabolism with serious consequences including mental retardation. The CBS domain appears to mediate regulation of activity of this protein by *S*-adenosylmethionine (Shan et al., 2001). As illustrated in Fig. 2, most IMPDH sequences (e.g. PIR-NREF: NF00078343 in superfamily SF000130) have two kinds of annotated Pfam domains, the catalytic IMP dehydrogenase/GMP reductase (IMPDH/GMPR) domain (PF00478), associated with PROSITE signature pattern (PS00487), and



```
ENTRY          T48678                                                       (A)
TITLE          proteasome alpha-1 chain [validated] - Haloferax volcanii
COMPLEX        heterodimer; alpha-1 and beta-1 (PIR:T48677) chain
               [validated; PMID:99412283]
FUNCTION       #description the predominant peptide-hydrolyzing activity of the
               alpha (1)beta(1)-proteasome is cleavage carboxyl to hydrophobic
               residues [validated; PMID:99412283]
```

```
ENTRY          XNHUSP  #type complete                                       (B)
TITLE          serine--pyruvate transaminase (EC 2.6.1.51), peroxisomal - human
FEATURE
  2-392        #product serine--pyruvate transaminase, peroxisomal
               #status experimental #label MAT\
   390-392     #region peroxisome/glyoxysome location signal #status atypical\
   2           #modified_site acetylated amino end (Ala) (in mature form)
               #status experimental [PMID:7798168]\
   209         #binding_site pyridoxal phosphate (Lys) (covalent) #status predicted\
   367         #binding_site carbohydrate (Asn) (covalent) #status absent
```

Fig. 1. PIR evidence attribution for: (A) title, complex, and function annotation; and (B) feature annotation.

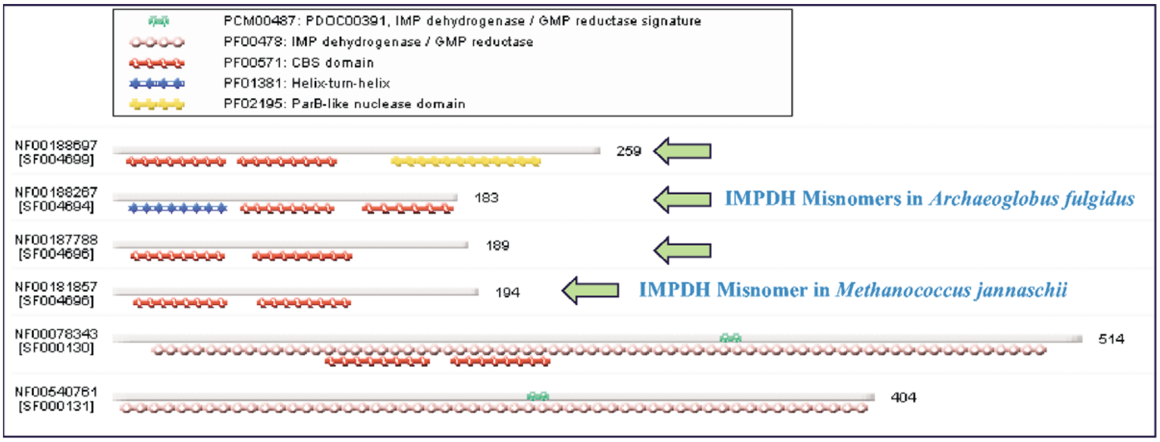

PCM00487: PDOC00391, IMP dehydrogenase / GMP reductase signature
PF00478: IMP dehydrogenase / GMP reductase
PF00571: CBS domain
PF01381: Helix-turn-helix
PF02195: ParB-like nuclease domain
NF00188697 [SF004699]
259
NF00188267 [SF004694]
183
IMPDH Misnomers in Archaeoglobus fulgidus
NF00187788 [SF004696]
189
NF00181857 [SF004696]
194
IMPDH Misnomer in Methanococcus jannaschii
NF00078343 [SF000130]
514
NF00540761 [SF000131]
404

Fig. 2. Domain architectures of IMP dehydrogenase (IMPDH) and misnomers. A typical IMPDH (NF00078343) has an IMPDH domain that forms the catalytic core and is interrupted by two CBS domains. A less common but functional IMPDH (NF00540761) lacks the CBS domains. All four misnomers show strong similarity to the CBS domains.

two adjacent CBS domains (PF00571), which actually interrupt the IMPDH/GMPR domain. Structurally, the N- and C-terminal parts of the IMPDH/GMPR domain form the core catalytic domain and the two CBS regions form a flanking globular domain (Zhang et al., 1999). There is also a well-characterized IMPDH (PIR-NREF: NF00540761 in SF000131) (Zhou et al., 1997) that contains the catalytic domain but lacks the CBS domains, showing that CBS domains are not necessary for enzymatic activity. The four misnomers shown in Fig. 2, one from the *Methanococcus jannaschii* genome and three from *Archaeoglobus fulgidus*, all lack the catalytic domain of IMPDH but contain adjacent CBS domains. Two of them also contain a domain usually associated with DNA binding (the ParB-like nuclease domain or the helix–turn–helix), which may provide a more reliable prediction for the functional classification of these proteins.

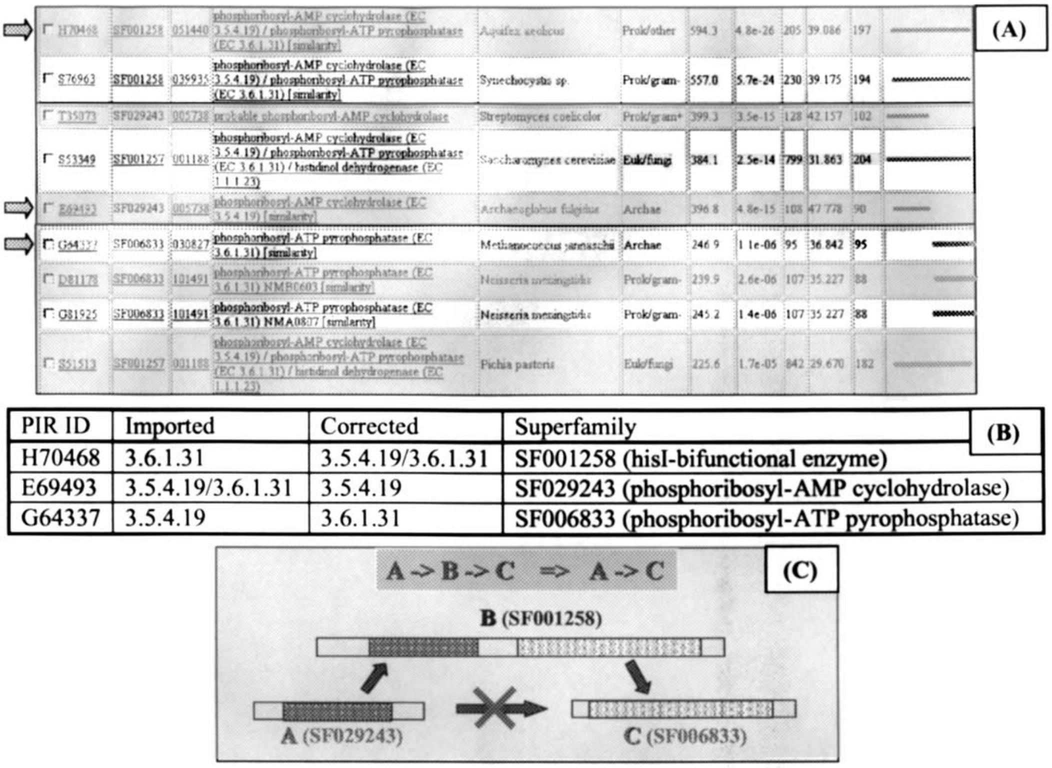
Many of the genome annotation errors still remain in sequence databases and have been propagated to secondary, curated databases. IMPDH occurs in most species, as the enzyme (EC 1.1.1.205) is the rate-limiting step in the de novo synthesis of guanine nucleotides. It is depicted in the Purine Metabolism pathway for *A. fulgidus* (afu00230) in the KEGG pathway database (Kanehisa et al., 2002) based on the three misannotated IMPDH proteins shown above. However, there is no evidence that a homologus IMPDH protein actually exists in the *A. fulgidus* genome to substantiate its placement on the pathway. Indeed, the only three proteins annotated by the genome center as IMPDH are all misnomers; and no IMPDH can be detected after genome-wide search using either sequence similarity searches (BLAST (Altschul et al., 1997) and/or FASTA (Pearson and Lipman, 1988)) against all known IMPDH proteins, or hidden Markov model search (HMMER (Eddy et al., 1995)) against the C-terminal part of the IMPDH/GMPR domain.

3.2. His-I bifunctional proteins: transitive identification catastrophe

Annotation errors originating from different genome centers have led to the so-called ‘transitive identification catastrophe.’ Fig. 3 illustrates an example where members of three related superfamilies were originally misannotated, likely because only local domain relationships were considered. Here, the related superfamilies are: SF001258, a bifunctional protein with two domains, for EC 3.5.4.19 and 3.6.1.31, respectively; SF029243, containing only the first domain, for EC 3.5.4.19; and SF006833, containing the second domain, for EC 3.6.1.31. Based on the superfamily classification, the improper names assigned to three sequence entries imported to PIR (H70468, E69493, G64337) were later corrected. The type of transitive annotation error observed in entry G64337 (named as EC 3.5.4.19 when it is actually EC 3.6.1.31) often involves multi-domain proteins. Comprehensive superfamily classification, thus, allows systematic detection and correction of genome annotation errors.

4. Analysis of the common identification errors

Faced with several thousands or tens of thousands of open reading frames to identify and functionally annotate, genome sequencing projects cannot be expected to perform a thorough examination of each molecule. For the most part, the sequence will be searched against a single comprehensive dataset, often NR (at NCBI (Wheeler et al., 2002)), PIR-PSD, or SwissProt/TrEMBL, and the sequence will be assigned the name



□ H70468	SF001258	051440	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Aquifex aeolicus	Prok/other	594.3	4.8e-26	205	39.086	197
□ S76963	SF001258	039935	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Synechocystis sp.	Prok/gram-	557.0	5.7e-24	230	39.175	194
□ T35073	SF029243	005738	probable phosphoribosyl-AMP cyclohydrolase	Streptomyces coelicolor	Prok/gram+	399.3	3.5e-15	128	42.157	102
□ S53349	SF001257	001188	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)	Saccharomyces cerevisiae	Euk/fungi	384.1	2.5e-14	799	31.863	204
□ E69493	SF029243	005738	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) [similarity]	Archaeoglobus fulgidus	Archae	396.8	4.8e-15	108	47.778	90
□ G64337	SF006833	030827	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]	Methanococcus jannaschii	Archae	246.9	1.1e-06	95	36.842	95
□ D81178	SF006833	101491	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMB0603 [similarity]	Neisseria meningitidis	Prok/gram-	239.9	2.6e-06	107	35.227	88
□ G81925	SF006833	101491	phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) NMA0837 [similarity]	Neisseria meningitidis	Prok/gram-	245.2	1.4e-06	107	35.227	88
□ S51513	SF001257	001188	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) / histidinol dehydrogenase (EC 1.1.1.23)	Pichia pastoris	Euk/fungi	225.6	1.7e-05	842	29.670	182

PIR ID	Imported	Corrected	Superfamily
H70468	3.6.1.31	3.5.4.19/3.6.1.31	SF001258 (hisI-bifunctional enzyme)
E69493	3.5.4.19/3.6.1.31	3.5.4.19	SF029243 (phosphoribosyl-AMP cyclohydrolase)
G64337	3.5.4.19	3.6.1.31	SF006833 (phosphoribosyl-ATP pyrophosphatase)

Fig. 3. Superfamily classification for correcting transitive identification error: (A) FASTA neighbors of H70468 are in three superfamilies; (B) misidentification of three proteins by genome centers was later corrected based on superfamily assignment; (C) the misannotation of G64337 is an example of transitive identification error.

of the highest-scoring sequence(s). Many database users also rely on searching a comprehensive database for the best-scoring retrieved matches in making identifications of unknown proteins. There are several problems with this approach. Firstly, the common sequence searching algorithms (BLAST, FASTA) find best-scoring similarities; however, the similarity may involve only parts of the query and target molecules, as illustrated by the numerous proteins misidentified as IMPDH. The retrieved similarity may be to a known domain that is tangential to the main function of the protein or to a region with compositional similarity, e.g. a region containing several transmembrane domains. Before making or accepting an identification, users should examine the domain structure in comparison to the pairwise alignments and determine if the similarity is local, perhaps associated with a common domain, or extends convincingly over the entire sequences.

Secondly, annotation in the searched databases is at best inconsistent and incomplete and at worst misleading or erroneous, having been based on partial or weak similarity. The major nucleotide sequence database GenBank/EMBL/DDBJ (Benson et al., 2002) is an 'archival' database, recording the original identifications as submitted by the sequencers unless a revision is submitted by the same group. Therefore, the protein identifications in GenPept, which are taken directly from GenBank annotations, may never be updated in light of more recent knowledge. Users need to realize that entries in a comprehensive database may be under-identified, e.g. labeled 'hypothetical protein' when there is a convincing similarity to a protein or domain of known function; over-identified, e.g. the specific activity 'trypsin' is ascribed when the less specific 'serine proteinase' would be more appropriate; or misidentified, as in the case studies discussed above.

Over-identification can be suspected when the similarity is not strong over the entire lengths of the query and target sequences. PIR defines 'closely related' as at least 50% identity (and with a significant e-value from FASTA search) and assigns such sequences to the same 'family'. A PIR superfamily is a collection of families. Sequences in different families in the same superfamily may have as little as 18–20% sequence identity and their activities, while often falling within the same general class, may be different. For example, the long-chain alcohol dehydrogenase superfamily contains alcohol dehydrogenase (EC 1.1.1.1), L-threonine 3-dehydrogenase (EC 1.1.1.103), L-iditol 2-dehydrogenase (EC 1.1.1.14), D-xylulose reductase (EC 1.1.1.9), galactitol-1-phosphate 5-dehydrogenase (EC 1.1.1.251), and others. Of five sequences from the recently sequenced genome of *Brucella melitensis* that were identified specifically as alcohol dehydrogenase (EC 1.1.1.1), only two are closely related (60% identity) to well-characterized alcohol dehydrogenases. For the others, the functional assignment may be overly specific, as they are more distantly related (less than 40% identity). For the most part, users will need to inspect database entries and read at least the abstracts of published reports to ascertain whether a functional assignment is based on experimental evidence or only on sequence similarity. Users should also ascertain that any residues critical for the ascribed activity (e.g. active site residues) are conserved.

Thirdly, in many cases a more thorough and time-consuming analysis is needed to reveal the most probable functional assignments. Factors that may be relevant, in addition to presence or absence of domains, motifs, or functional residues, include similarity or potential similarity of three-dimensional structures (when known), proximity of genes (may indicate that their products are involved in the same pathway), metabolic capacities of the organisms, and evolutionary history of the protein as deduced from aligned sequences. Bork and Koonin (1998) discuss additional effective strategies. Iyer et al. (2001) analyze several additional examples of misidentifications and their subsequent correction.

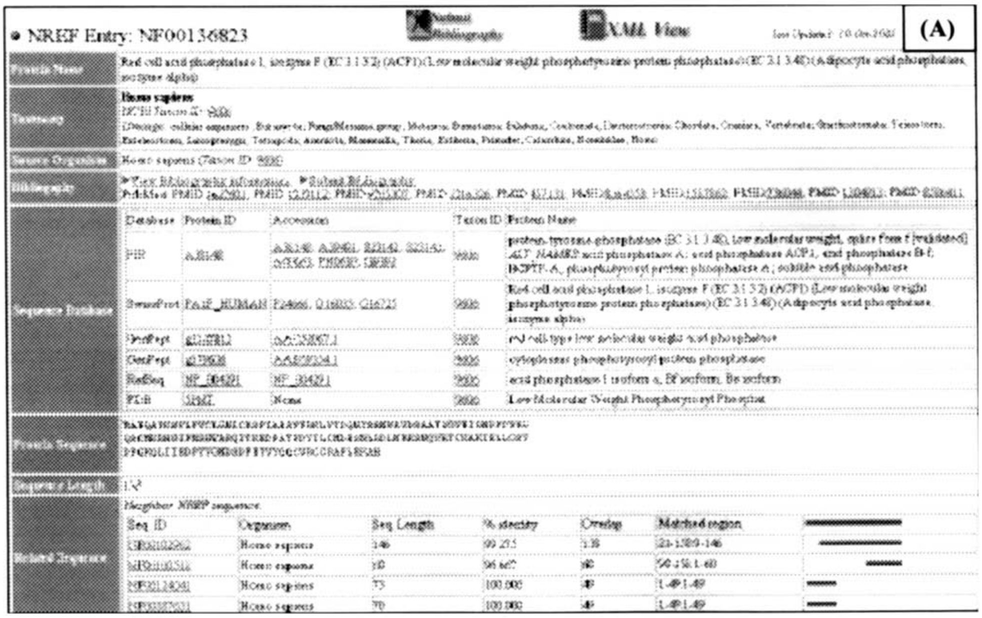
5. Integrated knowledge base system to facilitate functional annotation

To facilitate protein identification and functional annotation, two new protein databases (PIR-NREF and iProClass) have been developed and form a knowledge base system with sequence analysis tools and graphical user interfaces.

5.1. PIR-NREF non-redundant reference database

The PIR-NREF database provides a timely and comprehensive collection of protein sequence data containing source attribution and minimal redundancy. It consists of all sequences from PIR Protein Sequence Database, Swiss-Prot/TrEMBL, RefSeq, GenPept, and PDB (Westbrook et al., 2002), totaling more than 1,000,000 entries currently. Identical sequences from the same source organism (species) reported in different databases are presented as a single NREF entry with protein IDs, accession numbers, and protein names from each underlying database, as well as amino acid sequence, taxonomy, and composite bibliographic data (Fig. 4). Also listed are related sequences identified by all-against-all FASTA search, including identical sequences from different organisms, identical subsequences, and highly similar sequences (≥95% identity).

NREF can be used to assist functional identification of proteins, to develop an ontology of protein names, and to detect annotation errors. It is ideal for sequence analysis tasks because it is comprehensive, non-redundant, and contains composite annotations from source databases. The clustering at the species level aids analysis of evolutionary relationships of proteins. It also allows sequence searches against a subset of data consisting of sequences from one or more species. The collective protein names, including synonyms and alternate names, and the bibliographic information from all underlying databases provide an invaluable knowledge



• NREF Entry: NF00136823

last Update: [illegible]

Protein Name: Red cell acid phosphatase 1, isozyme F (EC 3.1.3.2) (ACP1) (Low molecular weight phosphotyrosine protein phosphatase) (EC 3.1.3.48) (Adipocyte acid phosphatase, isozyme alpha)

Taxonomy: Homo sapiens
NCBI Taxon ID: 9606
Lineage: cellular organisms; Eukaryota; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Primates; Catarrhini; Hominidae; Homo

Source Organism: Homo sapiens (Taxon ID: 9606)

Bibliography: ▶View Bibliographic information ▶Submit Bibliography
PubMed PMID [illegible]; PMID [illegible]; PMID [illegible]; PMID [illegible]; PMID [illegible]; PMID [illegible]; PMID [illegible]; PMID [illegible]; PMID [illegible]; PMID [illegible]

Sequence Database:

Database	Protein ID	Accession	Taxon ID	Protein Name
PIR	A38148	A38148, A39461, S23142, S23143, A35563, PH0689, S38382	9606	protein-tyrosine-phosphatase (EC 3.1.3.48), low molecular weight, splice form f [validated]; ACP; HAAP; acid phosphatase A; acid phosphatase ACP1; acid phosphatase B-f; HCPTP-A; phosphotyrosyl protein phosphatase A; soluble acid phosphatase
SwissProt	PPAC_HUMAN	P24666, Q16883, Q16725	9606	Red cell acid phosphatase 1, isozyme F (EC 3.1.3.2) (ACP1) (Low molecular weight phosphotyrosine protein phosphatase) (EC 3.1.3.48) (Adipocyte acid phosphatase, isozyme alpha)
GenPept	g[illegible]	AAC[illegible].1	9606	red cell-type low molecular weight acid phosphatase
GenPept	g[illegible]	AAA[illegible].1	9606	cytoplasmic phosphotyrosyl protein phosphatase
RefSeq	NP_004291	NP_004291.1	9606	acid phosphatase 1 isoform a, Bf isoform, Bs isoform
PDB	[illegible]	None	9606	Low Molecular Weight Phosphotyrosyl Phosphat

Protein Sequence:
MAEQATKSVLFVCLGNICRSPIAEAVFRKLVTDQNISENWRVDSAATSGYEIGNPPDYRG
QSCMKRHGIPMSHVARQITKEDFATFDYILCMDESNLRDLNRKSNQVKTCKAKIELLGSY
DPQKQLIIEDPYYGNDSDFETVYQQCVRCCRAFLEKAH

Sequence Length: 158

Related Sequence: Neighbor NREF sequence

Seq ID	Organism	Seq Length	% identity	Overlap	Matched region
NF[illegible]	Homo sapiens	146	99.275	138	21-158/9-146
NF[illegible]	Homo sapiens	60	96.667	60	99-158/1-60
NF[illegible]	Homo sapiens	73	100.000	49	1-49/1-49
NF[illegible]	Homo sapiens	70	100.000	49	1-49/1-49

Fig. 4. PIR-NREF sequence entry report. Each entry presents an identical sequence from the same source organism in one or more underlying protein databases.

base for application of natural language processing or computational linguistics techniques to develop a protein name ontology (Hirschman et al., in press). The different protein names assigned by different databases may also reflect annotation discrepancies. As an example, a protein (PIR: T40073) is variously named as a monofunctional (EC 3.5.4.19), bifunctional (EC 3.5.4.19, 3.6.1.31) or trifunctional (EC 3.5.4.19, 3.6.1.31, 1.1.1.23) protein in three different databases. Thus, the source name attribution provides clues to incorrectly annotated proteins.

5.2. iProClass integrated protein classification database

The iProClass database (Fig. 5) contains value-added descriptions of all proteins and serves as a framework for data integration in a distributed networking environment. It includes up-to-date information from many sources, thereby providing much richer annotation than can be found in any single database. The protein information in iProClass includes family relationships at both global (superfamily/family) and local (domain, motif, site) levels, as well as structural and functional classifications and features of proteins. The database is updated biweekly and currently consists of about 830,000 non-redundant protein sequences from the PIR-PSD, Swiss-Prot, and TrEMBL databases. The protein entries are organized with more than 36,000 PIR superfamilies, 145,000 families, 3700 Pfam and PIR homology domains, 1300 ProSite motifs, 550,000 FASTA similarity clusters, and links to over 50 molecular biology databases.

Database cross-references in iProClass are represented by rich links, which include both the links and related summary information. This approach effectively combines data warehouse and hypertext navigation methods for data integration to provide timely information from distributed sources. iProClass collects information from and links to databases for protein sequences (PIR-PSD, PIR-NREF, Swiss-Prot, TrEMBL, GenPept, RefSeq), families (InterPro, Pfam, ProSite, Blocks, Prints, COG, MetaFam, PIR-ASDB, ProClass), functions and pathways (EC-IUBMB, KEGG, BRENDA, WIT, MetaCyc, EcoCyc), interactions (DIP, BIND), post-translational modifications (RESID, PhosphoSite DB), protein expression and proteomes (PMG), structures and structural classifications (PDB, PDBSum, SCOP, CATH, FSSP, MMDB), genes and genomes (GenBank, EMBL, DDBJ, LocusLink, TIGR, SGD, FlyBase, MGI, GDB, OMIM, MIPS, GenProtEC), ontologies (GO), literature (PubMed), and taxonomy (NCBI Taxonomy).

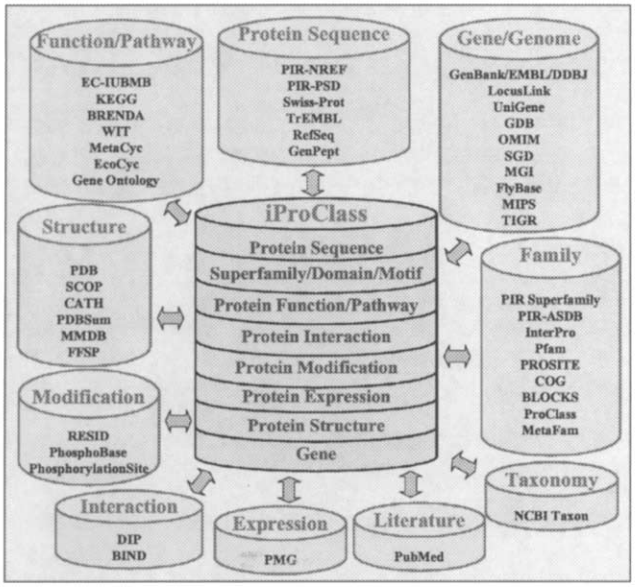
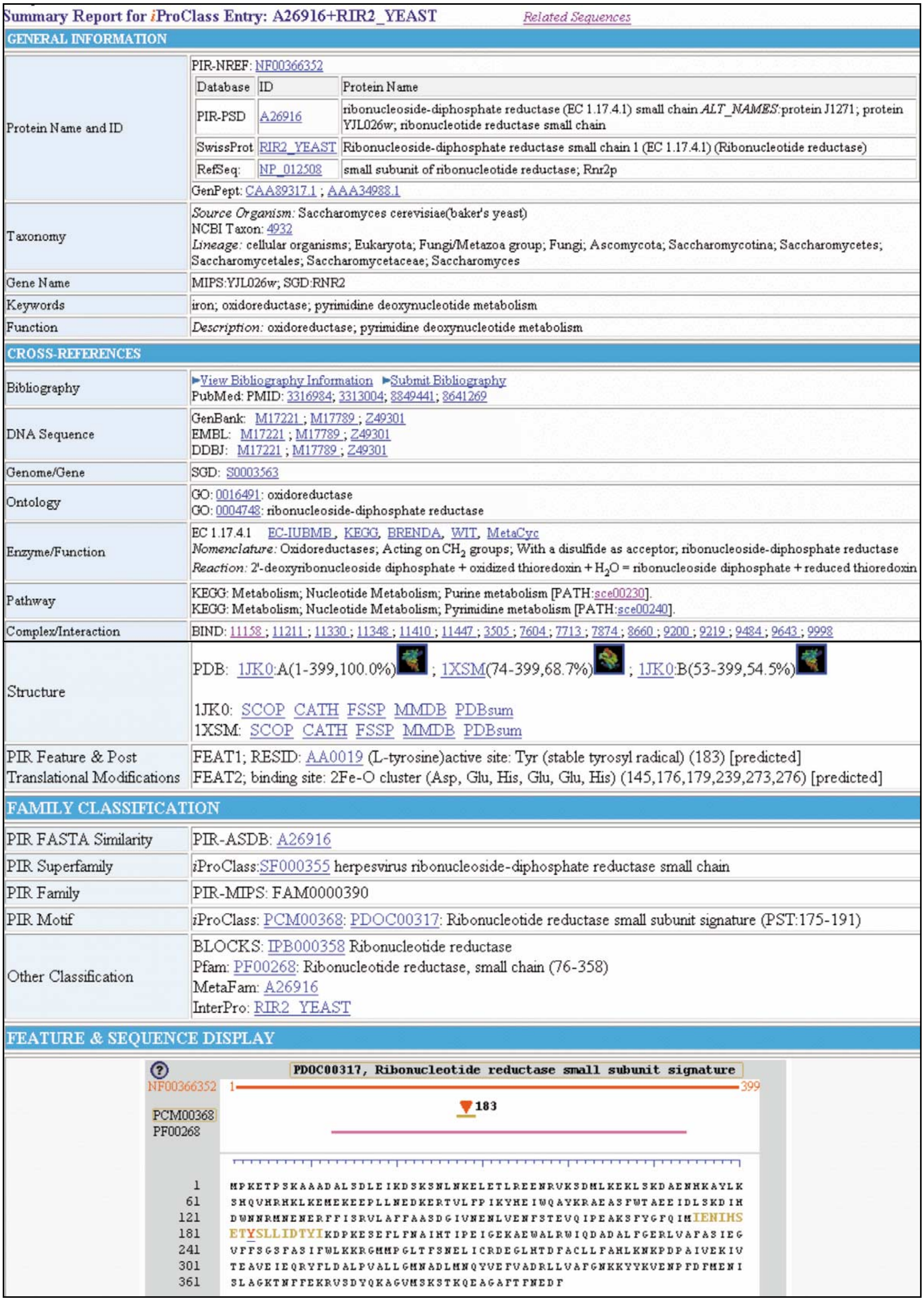


Fig. 5. iProClass database overview.

iProClass presents comprehensive views for protein sequences and superfamilies in two types of summary reports. The protein sequence report (Fig. 6) covers information on family, structure, function, gene, genetics, disease, ontology, taxonomy, and literature, with cross-references to relevant molecular databases and executive summary lines, as well as a graphical display of domain and motif sequence regions and a link to related sequences in pre-computed FASTA clusters. The superfamily report provides PIR superfamily membership information with length, taxonomy, and keyword statistics, complete member listing separated into major kingdoms, family relationships at the whole protein and domain and motif levels with direct mapping to other classifications, structure and function cross-references, graphical display of domain and motif architecture of members, and a link to dynamically generated multiple sequence alignments and phylogenetic trees for superfamilies with curated seed members.

5.3. Analytical tools and graphical interfaces

The PIR web site (http://pir.georgetown.edu) connects data mining and sequence analysis tools to underlying databases for information retrieval and knowledge discovery, with functionalities for interactive queries, combinations of sequence and annotation text searches, and sorting and visual exploration of search results. Direct entry report retrieval is based on sequence unique identifiers of all underlying databases, such as PIR, Swiss-Prot, or RefSeq. Basic and advanced text searches return protein entries listed in summary lines with information on protein IDs, matched fields, protein name, taxonomy, superfamily, domain, and motif, with hypertext links to the full entry report and to cross-referenced databases. More than 50 fields are searchable, including about 30 database unique identifiers (e.g. PDB ID, EC number, PubMed ID, and KEGG path-



Summary Report for *i*ProClass Entry: A26916+RIR2_YEAST *Related Sequences*

GENERAL INFORMATION

Protein Name and ID	PIR-NREF: NF00366352

Database	ID	Protein Name
PIR-PSD	A26916	ribonucleoside-diphosphate reductase (EC 1.17.4.1) small chain *ALT_NAMES:*protein J1271; protein YJL026w; ribonucleotide reductase small chain
SwissProt	RIR2_YEAST	Ribonucleoside-diphosphate reductase small chain 1 (EC 1.17.4.1) (Ribonucleotide reductase)
RefSeq:	NP_012508	small subunit of ribonucleotide reductase; Rnr2p

GenPept: CAA89317.1 ; AAA34988.1

Taxonomy	*Source Organism:* Saccharomyces cerevisiae(baker's yeast) NCBI Taxon: 4932 *Lineage:* cellular organisms; Eukaryota; Fungi/Metazoa group; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces
Gene Name	MIPS:YJL026w; SGD:RNR2
Keywords	iron; oxidoreductase; pyrimidine deoxynucleotide metabolism
Function	*Description:* oxidoreductase; pyrimidine deoxynucleotide metabolism

CROSS-REFERENCES

Bibliography	►View Bibliography Information ►Submit Bibliography PubMed: PMID: 3316984; 3313004; 8849441; 8641269
DNA Sequence	GenBank: M17221 ; M17789 ; Z49301 EMBL: M17221 ; M17789 ; Z49301 DDBJ: M17221 ; M17789 ; Z49301
Genome/Gene	SGD: S0003563
Ontology	GO: 0016491: oxidoreductase GO: 0004748: ribonucleoside-diphosphate reductase
Enzyme/Function	EC 1.17.4.1 EC-IUBMB , KEGG, BRENDA, WIT, MetaCyc *Nomenclature:* Oxidoreductases; Acting on CH_2 groups; With a disulfide as acceptor; ribonucleoside-diphosphate reductase *Reaction:* 2'-deoxyribonucleoside diphosphate + oxidized thioredoxin + H_2O = ribonucleoside diphosphate + reduced thioredoxin
Pathway	KEGG: Metabolism; Nucleotide Metabolism; Purine metabolism [PATH:sce00230]. KEGG: Metabolism; Nucleotide Metabolism; Pyrimidine metabolism [PATH:sce00240].
Complex/Interaction	BIND: 11158 ; 11211 ; 11330 ; 11348 ; 11410 ; 11447 ; 3505 ; 7604 ; 7713 ; 7874 ; 8660 ; 9200 ; 9219 ; 9484 ; 9643 ; 9998
Structure	PDB: 1JK0:A(1-399,100.0%) ; 1XSM(74-399,68.7%) ; 1JK0:B(53-399,54.5%) 1JK0: SCOP CATH FSSP MMDB PDBsum 1XSM: SCOP CATH FSSP MMDB PDBsum
PIR Feature & Post Translational Modifications	FEAT1; RESID: AA0019 (L-tyrosine)active site: Tyr (stable tyrosyl radical) (183) [predicted] FEAT2; binding site: 2Fe-O cluster (Asp, Glu, His, Glu, Glu, His) (145,176,179,239,273,276) [predicted]

FAMILY CLASSIFICATION

PIR FASTA Similarity	PIR-ASDB: A26916
PIR Superfamily	*i*ProClass:SF000355 herpesvirus ribonucleoside-diphosphate reductase small chain
PIR Family	PIR-MIPS: FAM0000390
PIR Motif	*i*ProClass: PCM00368: PDOC00317: Ribonucleotide reductase small subunit signature (PST:175-191)
Other Classification	BLOCKS: IPB000358 Ribonucleotide reductase Pfam: PF00268: Ribonucleotide reductase, small chain (76-358) MetaFam: A26916 InterPro: RIR2_YEAST

FEATURE & SEQUENCE DISPLAY

```
  1  MPKETPSKAAADALSDLEIKDSKSNLNKELETLREENRVKSDMLKEKLSKDAENHKAYLK
 61  SHQVHRHKLKEMEKEEPLLNEDKERTVLFPIKYHEIWQAYKRAEASFWTAEEIDLSKDIH
121  DWNNRMNENERFFISRVLAFFAASDGIVNENLVENFSTEVQIPEAKSFYGFQIMIENIHS
181  ETYSLLIDTYIKDPKESEFLFNAIHTIPEIGEKAEWALRWIQDADALFGERLVAFASIEG
241  VFFSGSFASIFWLKKRGMMPGLTFSNELICRDEGLHTDFACLLFAHLKNKPDPAIVEKIV
301  TEAVEIEQRYFLDALPVALLGMNADLMNQYVEFVADRLLVAFGNKKYYKVENPFDFMENI
361  SLAGKTNFFEKRVSDYQKAGVMSKSTKQEAGAFTFNEDF
```

Fig. 6. The iProClass sequence report for comprehensive value-added protein information.

way number) and a wide range of annotation texts (e.g. protein name, organism name, sequence feature, and paper title). The BLAST/FASTA search and peptide searches likewise return lists of matched entries with summary lines that also contain search statistics and matched sequence region. Protein entries returned from text and sequence searches can be selected for further analysis, including BLAST and FASTA search, pattern match, hidden Markov model (HMMER) domain search, ClustalW (Thompson et al., 1994) multiple sequence alignments, and PHYLIP (Felsenstein, 1989) phylogentic tree generation, and graphical display of superfamily, domain, and motif relationships. Species-based browsing and searching are supported for about 100 organisms, including over 70 complete genomes. Lists of related sequences in FASTA clusters can be retrievable, including sequence unique identifiers, annotation information, and graphical display of matched sequence regions.

6. Conclusion

The PIR serves as a primary resource for exploration of proteins, allowing users to answer complex biological questions that may typically involve querying multiple sources. In particular, interesting relationships between database objects, such as relationships among protein sequences, families, structures, and functions, can be revealed readily. Functional annotation of proteins requires association of proteins based on properties beyond sequence homology:proteins sharing common domains connected via related multi-domain proteins (grouped by superfamilies); proteins in the same pathways, networks, or complexes; proteins correlated in their expression patterns; and proteins correlated in their phylogenetic profiles (with similar evolutionary patterns) (Marcotte et al., 1999).

The data integration in PIR is important in revealing protein functional associations beyond sequence homology, as illustrated in the following example. As shown in Fig. 7A, the adenylylsulfate kinase (EC 2.7.1.25) domain (PF01583) appears in four different superfamilies (i.e. SF000544, SF001612, SF015480, SF003009), all having different overall domain arrangements. Except for SF000544, proteins in the other three superfamilies are bifunctional, all also containing sulfate adenylyltransferase (SAT) (EC 2.7.7.4) activity. However, the SAT enzymatic activity is found in two distinct sequence types, the ATP-sulfurylase (PF01747) domain and adjacent elongation factor Tu domains (PF00009 and PF03144), which share no detectable sequence similarity. Furthermore, both EC 2.7.1.25 and EC 2.7.7.4 are in adjacent steps of the same metabolic pathway (Fig. 7B). This example demonstrates that protein function may be revealed based on domain and/or pathway association, even without obvious sequence homology. The PIR knowledge base presents such complex superfamily–domain–function relationship to assist functional identification or characterization of proteins.

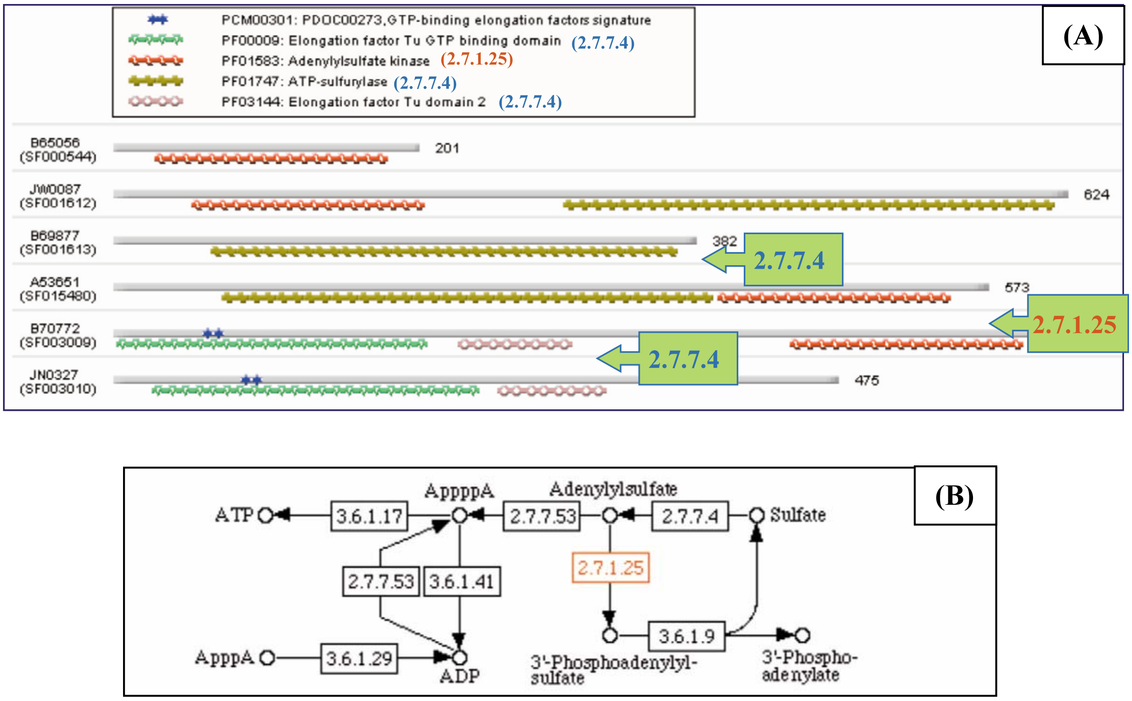

PCM00301: PDOC00273,GTP-binding elongation factors signature
PF00009: Elongation factor Tu GTP binding domain (2.7.7.4)
PF01583: Adenylylsulfate kinase (2.7.1.25)
PF01747: ATP-sulfurylase (2.7.7.4)
PF03144: Elongation factor Tu domain 2 (2.7.7.4)
(A)
B65056 (SF000544)
201
JW0087 (SF001612)
624
B69877 (SF001613)
382
2.7.7.4
A53651 (SF015480)
573
B70772 (SF003009)
2.7.1.25
2.7.7.4
JN0327 (SF003010)
475
(B)
ApppA
Adenylylsulfate
ATP
3.6.1.17
2.7.7.53
2.7.7.4
Sulfate
2.7.7.53
3.6.1.41
2.7.1.25
3.6.1.9
ApppA
3.6.1.29
ADP
3'-Phosphoadenylyl-sulfate
3'-Phospho-adenylate

Fig. 7. Superfamily–domain–function relationship for functional inference beyond sequence homology: (A) association of EC 2.7.1.25 and two distinct sequence types of EC 2.7.7.4 in multi-domain proteins; (B) association of EC 2.7.1.25 and EC 2.7.7.4 in the same metabolic pathway.

The PIR, with its integrated databases and analysis tools, thus constitutes a fundamental bioinformatics resource for biologists who contemplate using bioinformatics as an integral approach to their genomic/proteomic research and scientific inquiries.

Acknowledgements

The PIR is supported by grants P41 LM05978 and U01 HG02712 from National Institutes of Health. The iProClass database is supported by DBI-9974855 and DBI-0138188 from National Science Foundation.

References

Altschul, S.F., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Apweiler, R., et al., 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 29, 37–40.

Attwood, T.K., et al., 2002. PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res. 30, 239–241.

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 28, 45–48.

Barker, W.C., Pfeiffer, F., George, D.G., 1996. Superfamily classification in PIR-International Protein Sequence Database. Methods Enzymol. 266, 59–71.

Bateman, A., et al., 2002. The Pfam protein families database. Nucleic Acids Res. 30, 276–280.

Benson, D.A., et al., 2002. GenBank. Nucleic Acids Res. 30, 17–20.

Bork, P., Koonin, E.V., 1998. Predicting functions from protein sequences—where are the bottlenecks? Nat. Genet. 18, 313–318.

Brenner, S.E., 1999. Errors in genome annotation. Trends Genet. 15, 132–133.

Corpet, F., Servant, F., Gouzy, J., Kahn, D., 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. 28, 267–269.

Dayhoff, M.O., 1965–1978. Atlas of Protein Sequence and Structure. vol. 1–5, supplements 1–3. National Biomedical Research Foundation, Washington, DC.

Dayhoff, M.O., 1976. The origin and evolution of protein superfamilies. Fed. Proc. 35, 2132–2138.

Devos, D., Valencia, A., 2001. Intrinsic errors in genome annotation. Trends Genet. 17, 429–431.

Eddy, S.R., Mitchison, G., Durbin, R., 1995. Maximum discrimination hidden Markov models of sequence consensus. J. Comp. Biol. 2, 9–23.

Falquet, L., et al., 2002. The PROSITE database, its status in 2002. Nucleic Acids Res. 30, 235–238.

Felsenstein, J., 1989. PHYLIP—Phylogeny Inference Package (version 3.2). Cladistics 5, 164–166.

Haft, D.H., et al., 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. 29, 41–43.

Hirschman, L., Park, J.C., Tsujii, J., Wong, L., Wu, C.H., 2002. Accomplishments and challenges in literature data mining for Biology. Bioinformatics 18 (in press).

Huang, H., Barker, W.C., Chen, Y., Wu, C.H., 2003. iProClass: an integrated database of protein family, function, and structure information. Nucleic Acids Res. 31 (in press).

Iyer, L.M. et al., 2001. Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol. 2 (12), research0051.1–0051.11.

Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A., 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30, 42–46.

Lo Conte, L., Brenner, S.E., Hubbard, T.J.P, Chothia, C., Murzin, A.G., 2002. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res. 30, 264–267.

Marcotte, E.M, Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D., 1999. A combined algorithm for genome-wide prediction of protein function. Nature 402, 83–86.

Pearl, F.M.G., et al., 2001. A rapid classification protocol for the CATH domain database to support structural genomics. Nucleic Acids Res. 29, 223–227.

Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85, 2444–2448.

Pruitt, K.D., Maglott, D.R., 2001. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 29, 137–140.

Shan, X., Dunbrack, R.L., Jr., Christopher, S.A., Kruger, W.D., 2001. Mutations in the regulatory domain of cystathionine beta synthase can functionally suppress patient-derived mutations in cis. Hum. Mol. Genet. 10, 635–643.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Westbrook, J., et al., 2002. The Protein Data Bank: unifying the archive. Nucleic Acids Res. 30, 245–248.

Wheeler, D.L., et al., 2002. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res. 30, 13–16.

Wu, C.H. et al., 2003. The Protein Information Resource. Nucleic Acids Res. 31 (in press).

Yona, G., Linial, N., Linial, M., 2000. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res. 28, 49–55.

Zhang, R., et al., 1999. Characteristics and crystal structure of bacterial inosine-5′-monophosphate dehydrogenase. Biochemistry 38, 4691–4700.

Zhou, X., Cahoon, M., Rosa, P., Hedstrom, L., 1997. Expression, purification, and characterization of inosine 5′-monophosphate dehydrogenase from Borrelia burgdorferi. J. Biol. Chem. 272, 21977–21981.