CHAPTER 19

A FAMILY CLASSIFICATION APPROACH TO FUNCTIONAL ANNOTATION OF PROTEINS

Cathy H. Wu

Georgetown University Medical Center
wuc@georgetown.edu

Winona C. Barker

Georgetown University Medical Center
wb8@georgetown.edu

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. While this is a powerful approach that has led to many scientific discoveries, accurate annotation often requires the use of a variety of algorithms and databases, coupled with manual curation. This complex and ambiguous process is inevitably error prone.[92] Indeed, numerous genome annotation errors have been detected,[104, 203] 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, is itself incorrectly predicted, or simply has a different function.

The Protein Information Resource (PIR)[903] provides an integrated public resource of protein informatics to support genomic and proteomic research and scientific discovery. PIR produces the Protein Sequence Database (PSD) of functionally annotated protein sequences, which grew out of the Atlas of Protein Sequence and Structure edited by Margaret Dayhoff.[191] The annotation problems are addressed by a classification-driven and rule-based method with evidence at-

tribution, 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 chapter describes and illustrates how to use PIR databases and tools for functional annotation of proteins with case studies.

ORGANIZATION.

Section 1. We present a detail description of the classification-driver rule-based approach in PIR to the functional annotation of proteins.

Section 2. Then we illustrate the approach by two case studies. The first case study looks at the issue of error propagation to secondary databases using the example of IMP Dehydrogenase. The second case study looks at the issue of transitive identification error using the example of His-I bifunctional proteins.

Section 3. After that, we provide a careful discussion on the common identification errors and their causes.

Sections 4–5. Finally, we describe two new protein databases (NREF and iProClass) of PIR in details. We also discuss how this integrated knowledge base can facilitate protein function annotation in a way that goes beyond sequence homology.

1. Classification-Driven and Rule-Based Annotation with Evidence Attribution

1.1. *Protein Family Classification*

Classification of proteins is widely accepted to provide valuable clues to structure, activity, and metabolic role. This is increasingly important in this era of complete genome sequencing. Protein family classification has several advantages as a basic approach for large-scale 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[62] in the PIR-PSD, and protein groups in ProtoMap;[919]
(2) families of protein domains, such as those in Pfam[64] and ProDom;[175]
(3) sequence motifs or conserved regions, such as in PROSITE[238] and PRINTS;[37]
(4) structural classes, such as in SCOP[517] and CATH;[651] as well as
(5) integrations of various family classifications, like ProClass/iProClass[370, 904] and InterPro.[29]

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,[193] the original such classification based on sequence similarity, is unique in providing comprehensive and non-overlapping clustering of sequences into a hierarchical ordering of proteins to reflect their evolutionary origins and relationships. Proteins are assigned to the same superfamily/family only if they share end-to-end sequence similarity, including similar 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.[313] 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.

1.2. *Rule-Based Annotation and Evidence Attribution*

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 system and sequence patterns or profiles, numerous rules have been defined to predict positionspecific 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 the feature "Binding site: 2Fe-2S cluster (Cys) (covalent)" is added if the pattern is found.

Such sequence features are most accurately predicted if based on patterns or profiles derived from sequences most 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, misspelt enzyme names, or inconsistent EC number and enzyme name.

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. The entries are also tagged with "experimental," "predicted," "absent," or "atypical" in Feature annotations.

The validated Function or Complex annotation includes hypertext-linked PubMed unique identifiers for the articles in which the experimental determinations are reported. 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.

2. Case Studies

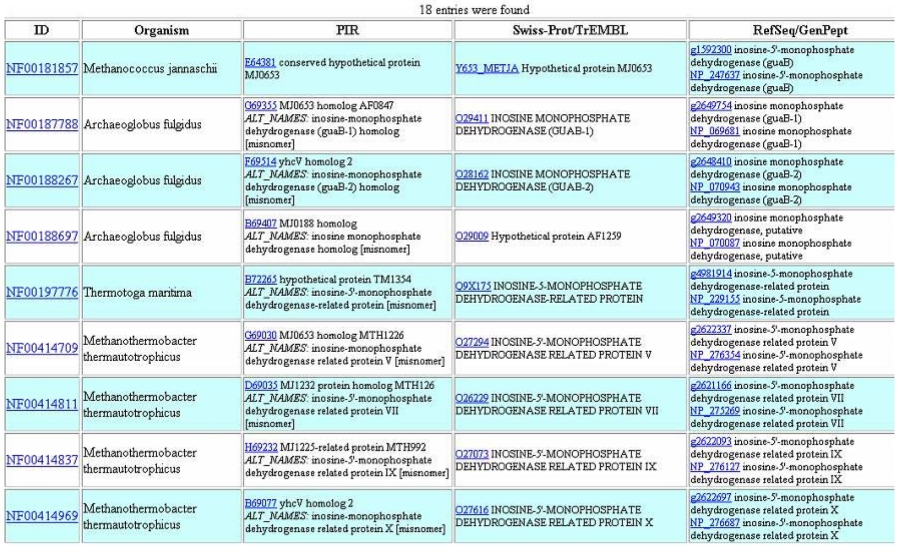
2.1. *IMP Dehydrogenase: Error Propagation to Secondary Databases*

During the PIR superfamily classification and curation process, at least 18 proteins were found to be mis-annotated 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,[41] still exist in GenPept (annotated GenBank translations) and RefSeq[684], see Figure 1.

The mis-annotation apparently resulted from local sequence similarity to the CBS domain. As illustrated in Figure 2, most IMPDH sequences (*e.g.*, PIR: A31997 in superfamily SF000130) have four annotated Pfam domains, the N-terminal IMPDH/GMP reductase domain (PF01574), the C-terminal IMPDH/GMP reductase domain (PF00478) associated with a PROSITE signature pattern (PS00487), and two CBS domains (PF00571). [Note added in press: PF01574 and PF00478 are now represented by one single Pfam domain PF00478.] Structurally, the N- and C-terminal domains form the core catalytic domain and the two CBS regions form a flanking CBS dimer domain.[932] There is also a well-characterized IMPDH (PIR: E70218 in SF000131)[938] that contains the N- and C-terminal catalytic domains but lacks the CBS domains, showing that CBS domains are not necessary for enzymatic activity.

The four misnomers shown in Figure 2, one from the *Methanococcus jannaschii* genome and three from *Archaeoglobus fulgidus*, all lack the functional region of an IMPDH but contain the two repeating CBS domains. Two of them also possess other domains, and have been classified into different superfamilies.

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 *Archaeoglobus fulgidus* (afu00230) in the KEGG pathway database[410] based on the three mis-annotated IMPDH proteins shown above. However, there is no evidence that a homologous IMPDH protein actually exists in the *Archaeoglobus 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[24] and/or FASTA[655]) against all known IMPDH proteins, or hidden Markov model search (HMMER[221]) against the N- and C-terminal IMPDH domains.



18 entries were found

ID	Organism	PIR	Swiss-Prot/TrEMBL	RefSeq/GenPept
NF00181857	Methanococcus jannaschii	E64381 conserved hypothetical protein MJ0653	Y653_METJA Hypothetical protein MJ0653	g1592300 inosine-5'-monophosphate dehydrogenase (guaB) NP_247637 inosine-5'-monophosphate dehydrogenase (guaB)
NF00187788	Archaeoglobus fulgidus	G69355 MJ0653 homolog AF0847 *ALT_NAMES*: inosine-monophosphate dehydrogenase (guaB-1) homolog [misnomer]	O29411 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-1)	g2649754 inosine monophosphate dehydrogenase (guaB-1) NP_069681 inosine monophosphate dehydrogenase (guaB-1)
NF00188267	Archaeoglobus fulgidus	F69514 yhcV homolog 2 *ALT_NAMES*: inosine-monophosphate dehydrogenase (guaB-2) homolog [misnomer]	O28162 INOSINE MONOPHOSPHATE DEHYDROGENASE (GUAB-2)	g2648410 inosine monophosphate dehydrogenase (guaB-2) NP_070943 inosine monophosphate dehydrogenase (guaB-2)
NF00188697	Archaeoglobus fulgidus	B69407 MJ0188 homolog *ALT_NAMES*: inosine monophosphate dehydrogenase homolog [misnomer]	O29009 Hypothetical protein AF1259	g2649320 inosine monophosphate dehydrogenase, putative NP_070087 inosine monophosphate dehydrogenase, putative
NF00197776	Thermotoga maritima	B72265 hypothetical protein TM1354 *ALT_NAMES*: inosine-5'-monophosphate dehydrogenase-related protein [misnomer]	Q9X175 INOSINE-5-MONOPHOSPHATE DEHYDROGENASE-RELATED PROTEIN	g4981914 inosine-5-monophosphate dehydrogenase-related protein NP_229155 inosine-5-monophosphate dehydrogenase-related protein
NF00414709	Methanothermobacter thermautotrophicus	G69030 MJ0653 homolog MTH1226 *ALT_NAMES*: inosine-monophosphate dehydrogenase related protein V [misnomer]	O27294 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN V	g2622337 inosine-5'-monophosphate dehydrogenase related protein V NP_276354 inosine-5'-monophosphate dehydrogenase related protein V
NF00414811	Methanothermobacter thermautotrophicus	D69035 MJ1232 protein homolog MTH126 *ALT_NAMES*: inosine-5'-monophosphate dehydrogenase related protein VII [misnomer]	O26229 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN VII	g2621166 inosine-5'-monophosphate dehydrogenase related protein VII NP_275269 inosine-5'-monophosphate dehydrogenase related protein VII
NF00414837	Methanothermobacter thermautotrophicus	H69232 MJ1225-related protein MTH992 *ALT_NAMES*: inosine-5'-monophosphate dehydrogenase related protein IX [misnomer]	O27073 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN IX	g2622093 inosine-5'-monophosphate dehydrogenase related protein IX NP_276127 inosine-5'-monophosphate dehydrogenase related protein IX
NF00414969	Methanothermobacter thermautotrophicus	B69077 yhcV homolog 2 *ALT_NAMES*: inosine-monophosphate dehydrogenase related protein X [misnomer]	O27616 INOSINE-5'-MONOPHOSPHATE DEHYDROGENASE RELATED PROTEIN X	g2622697 inosine-5'-monophosphate dehydrogenase related protein X NP_276687 inosine-5'-monophosphate dehydrogenase related protein X

Fig. 1. A partial list of the 18 IMP dehydrogenase misnomers in complete genomes remaining in some protein databases.

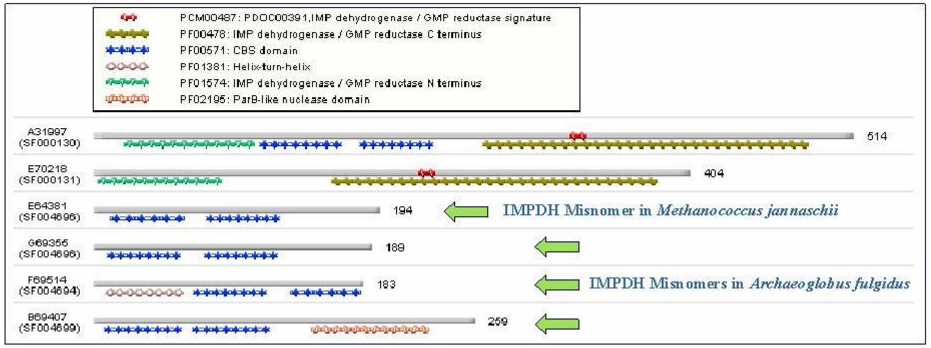

PCM00487: PDOC00391,IMP dehydrogenase / GMP reductase signature
PF00478: IMP dehydrogenase / GMP reductase C terminus
PF00571: CBS domain
PF01381: Helix-turn-helix
PF01574: IMP dehydrogenase / GMP reductase N terminus
PF02195: ParB-like nuclease domain
A31997 (SF000130)
514
E70218 (SF000131)
404
E64381 (SF004696)
194
IMPDH Misnomer in *Methanococcus jannaschii*
G69355 (SF004696)
189
F69514 (SF004694)
183
IMPDH Misnomers in *Archaeoglobus fulgidus*
B69407 (SF004699)
259

Fig. 2. Domain architectures of IMP dehydrogenase (IMPDH) and misnomers. A typical IMPDH (A31997) has two IMPDH domains that form the catalytic core and two CBS domains. A less common but functional IMPDH (E70218) lacks the CBS domains. All four misnomers show strong similarity to the CBS domains.

2.2. *His-I Bifunctional Proteins: Transitive Identification Catastrophe*

Several annotation errors originating from different genome centers have led to the so-called "transitive identification catastrophe." Figures 3 and 4 illustrate an example where members of three related superfamilies were originally mis-annotated, 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.

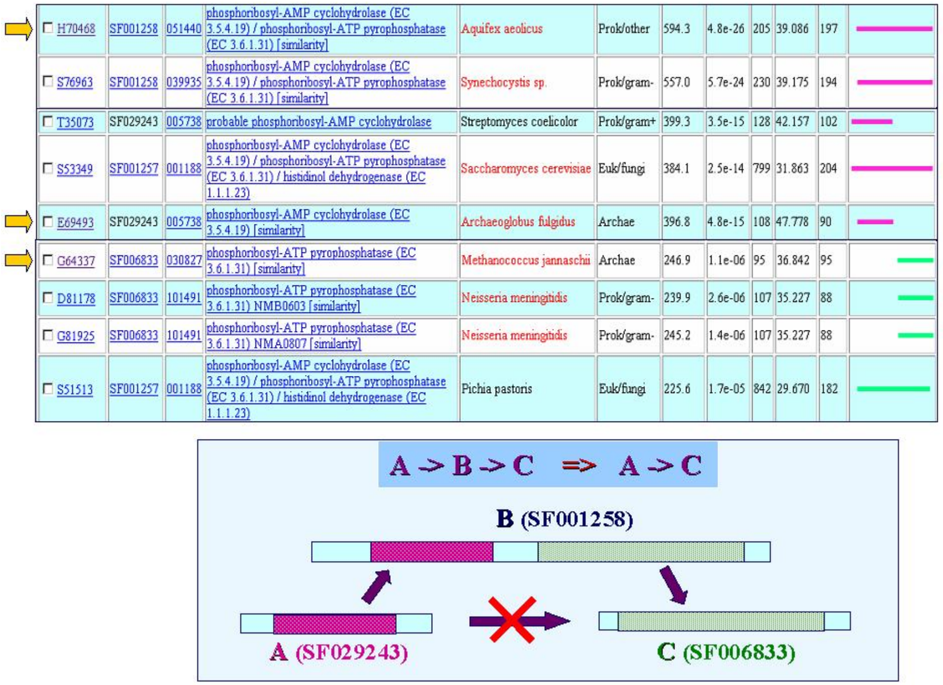
Based on the superfamily classification, the improper names assigned to three sequence entries imported to PIR (H70468, E69493, G64337) were later corrected. The type of transitive annotation error observed in entry G64337 (named as EC 3.5.4.19 when it is actually EC 3.6.1.31) often involves multi-domain proteins. Comprehensive superfamily and domain classification, thus, allows systematic detection and correction of genome annotation errors.

3. 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[882], PIR-PSD, or SwissProt/TrEMBL, and the sequence will be assigned the name 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 mis-identified 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



⇨	□ 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) NMA0807 [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	▬

Fig. 3. FASTA neighbors of H70468 are in three superfamilies. G64337 is an example of mis-annotation by transitive identification error. It was named as EC 3.5.4.19 when it is actually EC 3.6.1.31.

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 pyeophosphatase

Fig. 4. The mis-identification of three proteins by genome centers was later corrected based on the superfamily assignments in Figure 3.

and incomplete and at worst misleading or erroneous, having been based on partial or weak similarity. The major nucleotide sequence database GenBank/EMBL/DDBJ[77] 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 mis-identified, 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 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 [92] discuss additional effective strategies. Iyer *et al.*[387] analyze several additional examples of mis-identifications and their subsequent correction.

4. Integrated Knowledge Base System to Facilitate Functional Annotation

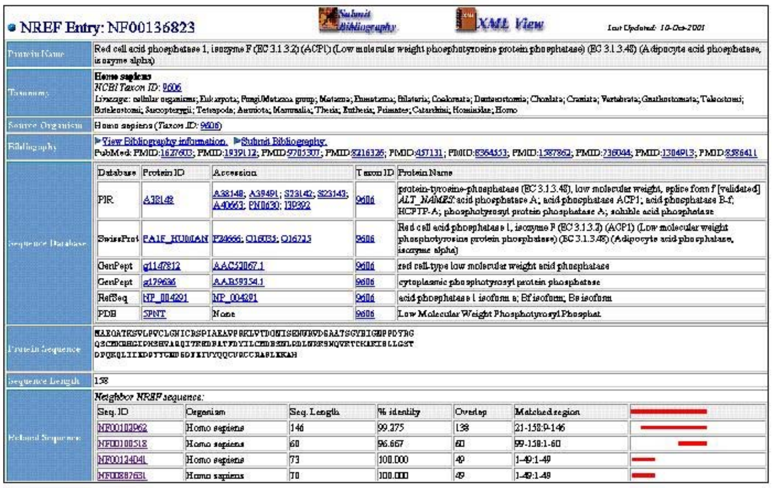
To facilitate protein identification and functional annotation, two new protein databases (NREF and iProClass) have been developed and are being integrated into a knowledge base system with sequence analysis tools and graphical user interfaces.

4.1. *PIR-NREF Non-Redundant Reference Database*

The PIR-NREF database was designed to provide all the identifications in major databases for any given sequence, identified by source organisms. It is a timely and comprehensive collection of all protein sequence data containing source attribution and minimal redundancy. The database has three major features:

(1) comprehensiveness and timeliness: it currently consists of more than 920,000 sequences from PIR Protein Sequence Database, Swiss-Prot/TrEMBL, RefSeq, GenPept, and PDB,[880] and is updated biweekly;
(2) non-redundancy: it is clustered by sequence identity and taxonomy at the species level; and
(3) source attribution: it contains protein IDs, accession numbers, and protein names from source databases in addition to amino acid sequence, taxonomy,

(I)



NREF Entry: NF00136823 Submit Bibliography XML View Last Updated: 10-Oct-2001

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 Bibliography information. Submit Bibliography. PubMed: PMID:1627603; PMID:1939112; PMID:9705307; PMID:8216326; PMID:457131; PMID:8364553; PMID:1387862; PMID:736044; PMID:1304913; PMID:8586411

Sequence Database

Database	Protein ID	Accession	Taxon ID	Protein Name
PIR	A38148	A38148; A39491; S23142; S23143; A40663; PN0630; I39392	9606	protein-tyrosine-phosphatase (EC 3.1.3.48), low molecular weight, splice form f [validated] *ALT_NAMES:* acid phosphatase A; acid phosphatase ACP1; acid phosphatase B-f; HCPTP-A; phosphotyrosyl protein phosphatase A; soluble acid phosphatase
SwissProt	PA1F_HUMAN	P24666; Q16035; Q16713	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	g1147812	AAC52067.1	9606	red cell-type low molecular weight acid phosphatase
GenPept	g179636	AAB59354.1	9606	cytoplasmic phosphotyrosyl protein phosphatase
RefSeq	NP_004291	NP_004291	9606	acid phosphatase 1 isoform a; Bf isoform; Bs isoform
PDB	5PNT	None	9606	Low Molecular Weight Phosphotyrosyl Phosphat

Protein Sequence

```
MAEQATKSVLFVCLGNICRSPIAEAVFRKLVTDQNISENWRVDSAATSGYEIGNPPDYRG
QGCMKRHGIPMSHVARQITKEDFATFDYILCMDESNLRDLNRKSNQVKTCKAKIELLGSY
DPQKQLIIEDPYYGNDSDFETVYQQCVRCCRAFLEKAH
```

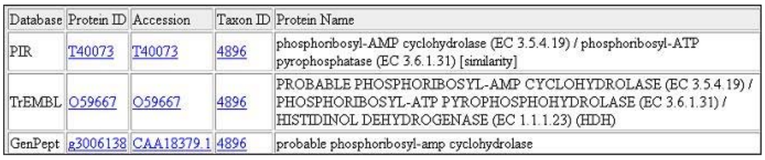
Sequence Length 158

Related Sequences

Neighbor NREF sequence:

Seq. ID	Organism	Seq. Length	% identity	Overlap	Matched region
NF00102962	Homo sapiens	146	99.275	138	21-158:9-146
NF00100518	Homo sapiens	60	96.667	60	99-158:1-60
NF00124041	Homo sapiens	73	100.000	49	1-49:1-49
NF00087631	Homo sapiens	70	100.000	49	1-49:1-49

(II)



Database	Protein ID	Accession	Taxon ID	Protein Name
PIR	T40073	T40073	4896	phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19) / phosphoribosyl-ATP pyrophosphatase (EC 3.6.1.31) [similarity]
TrEMBL	O59667	O59667	4896	PROBABLE PHOSPHORIBOSYL-AMP CYCLOHYDROLASE (EC 3.5.4.19) / PHOSPHORIBOSYL-ATP PYROPHOSPHOHYDROLASE (EC 3.6.1.31) / HISTIDINOL DEHYDROGENASE (EC 1.1.1.23) (HDH)
GenPept	g3006138	CAA18379.1	4896	probable phosphoribosyl-amp cyclohydrolase

Fig. 5. (I) PIR-NREF sequence entry report. Each entry presents an identical sequence from the same source organism in one or more underlying protein databases. (II) Discrepant protein names assigned by different databases reveal annotation errors.

and composite bibliographic data (Figure 5, Part I). Related sequences, including identical sequences from different organisms, as well as identical subsequences and highly similar sequences ($\geq 95\%$ sequence identity) are also listed. The NCBI taxonomy[882] is used for matching source organism names at the species or strain (if known) levels.

The PIR-NREF database 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 composite protein names, including synonyms and alternate names, and the bibliographic information from all underlying databases provide an invaluable knowledge base for application of natural language processing or computational linguistics techniques to develop a protein name ontology.[355, 920]

The different protein names assigned by different databases may also reflect annotation discrepancies. As an example (Figure 5, Part II), the 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.

4.2. *iProClass Integrated Protein Classification Database*

A few sentences describing the properties of a protein may not be adequate annotation. What is required is as much reliable information as possible about properties like function(s) of the protein, domains and sites, catalytic activity, pathways, subcellular location, processes in which the protein may be involved, similarities to other proteins, *etc.* Thus, an ideally annotated protein database should

(1) include domain structure, motif identification, and classifications,
(2) distinguish experimentally determined from predicted information, with citations for the former and method for the latter, and
(3) include annotations of gene location, expression, protein interactions, and structure determinations.

However, in practice, it is unrealistic to expect that protein sequence databases can keep all (or even a substantial minority) of entries up-to-date with regard to all of the above. Nevertheless, much of this information is available in specialty databases.

The iProClass database was designed to include up-to-date information from many sources, thereby, providing much richer annotation than can be found in any single database. It contains value-added descriptions of all proteins and serves as a framework for data integration in a distributed networking environment. 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 extended from ProClass,[370, 902] a protein family database that organizes proteins based on PIR superfamilies and PROSITE motifs.

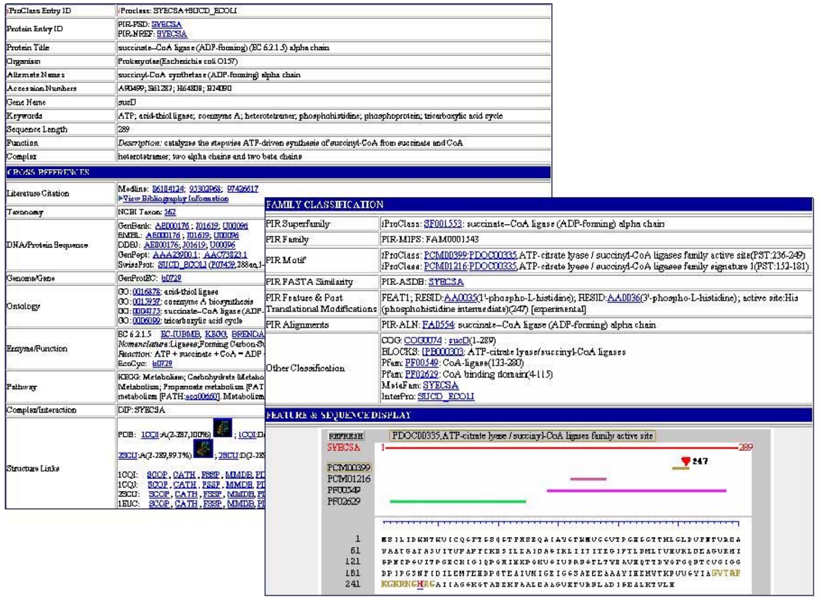
The version at the time of writing (May 2002) consists of more than 735,000 non-redundant PIR-PSD, SwissProt, and TrEMBL proteins organized with more than 36,000 PIR superfamilies, 145,000 families, 3700 Pfam and PIR homology domains, 1300 PROSITE/ProClass motifs, 280 RESID[270] post-translational modification sites, 550,000 FASTA similarity clusters, and links to over 45 molecular biology databases. iProClass cross-references include databases for protein families (*e.g.*, COG[819] and InterPro), functions and pathways (*e.g.*, KEGG[410] and WIT[642]), interactions (*e.g.*, DIP[910]), structures and structural classifications (*e.g.*, PDB, SCOP, CATH, and PDBSum[475]), genes and genomes (*e.g.*, TIGR[666] and OMIM[318]), ontologies (*e.g.*, Gene Ontology[285]), literature (*e.g.*, NCBI PubMed), and taxonomy (*e.g.*, NCBI Taxonomy).

The iProClass presents comprehensive protein and superfamily views as sequence and superfamily summary reports. The protein sequence report (Figure 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 regions. The superfamily report (Figure 7) 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, and domain and motif graphical display.

4.3. *Analytical Tools and Graphical Interfaces*

Integrated with the protein databases are many search and analysis tools that are freely accessible from the PIR website (`http://pir.georgetown.edu`)[555]. These tools assist the exploration of protein structure and function for knowledge discovery. For reliable protein identification, search results should display more detailed information, including lengths of the query and target sequences and of the "overlap" (the local regions of each that are matched), and the percentage identity of the overlap region, as in the interface that displays FASTA neighbors (Figure 3).

Other useful features available from the PIR website are graphical displays, such as the domain and motif display in iProClass reports (Figures 6 and 7); ability to make multiple alignments and display domain structures; ability to sort search output by criteria (*e.g.*, species) other than similarity score; and easy retrieval of full entries, citation abstracts, and classification information with multiple text search options.



iProClass Entry ID	iProclass: SYECSA+SUCD_ECOLI
Protein Entry ID	PIR-PSD: SYECSA PIR-NREF: SYECSA
Protein Title	succinate–CoA ligase (ADP-forming) (EC 6.2.1.5) alpha chain
Organism	Prokaryotae(Escherichia coli O157)
Alternate Names	succinyl-CoA synthetase (ADP-forming) alpha chain
Accession Numbers	A90499; S61287; H64808; B24090
Gene Name	sucD
Keywords	ATP; acid-thiol ligase; coenzyme A; heterotetramer; phosphohistidine; phosphoprotein; tricarboxylic acid cycle
Sequence Length	289
Function	*Description:* catalyzes the stepwise ATP-driven synthesis of succinyl-CoA from succinate and CoA
Complex	heterotetramer; two alpha chains and two beta chains
CROSS REFERENCES	
Literature Citation	Medline: 86104124; 93302968; 97426617 ▶View Bibliography Information
Taxonomy	NCBI Taxon: 562
DNA/Protein Sequence	GenBank: AE000176 ; J01619; U00096 EMBL: AE000176 ; J01619; U00096 DDBJ: AE000176; J01619; U00096 GenPept: AAA23900.1; AAC73823.1 SwissProt: SUCD_ECOLI (P07459,288aa,1-
Genome/Gene	GenProtEC: b0729
Ontology	GO : 0016878: acid-thiol ligase GO : 0015937: coenzyme A biosynthesis GO : 0004775: succinate–CoA ligase (ADP- GO : 0006099: tricarboxylic acid cycle
Enzyme/Function	EC 6.2.1.5 EC-IUBMB, KEGG, BRENDA *Nomenclature*:Ligases;Forming Carbon-Su *Reaction*: ATP + succinate + CoA = ADP EcoCyc: b0729
Pathway	KEGG: Metabolism; Carbohydrate Metabo Metabolism; Propanoate metabolism [PAT metabolism [PATH:eco00660]. Metabolism
Complex/Interaction	DIP: SYECSA
Structure Links	PDB: 1CQI:A(2-287,100%) ; 1CQI:D 2SCU:A(2-289,99.7%) ; 2SCU:D(2-28 1CQI: SCOP , CATH , FSSP , MMDB, PD 1CQJ: SCOP , CATH , FSSP , MMDB, PD 2SCU: SCOP , CATH , FSSP , MMDB, PD 1EUC: SCOP , CATH , FSSP , MMDB, PD

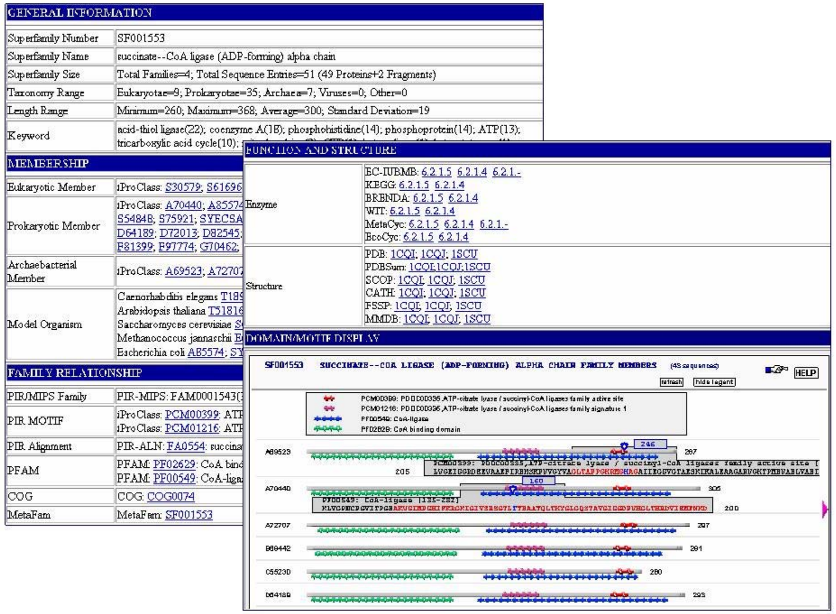
FAMILY CLASSIFICATION	
PIR Superfamily	iProClass: SF001553: succinate–CoA ligase (ADP-forming) alpha chain
PIR Family	PIR-MIPS: FAM0001543
PIR Motif	iProClass: PCM00399:PDOC00335,ATP-citrate lyase / succinyl-CoA ligases family active site(PST:236-249) iProClass: PCM01216:PDOC00335,ATP-citrate lyase / succinyl-CoA ligases family signature 1(PST:152-181)
PIR FASTA Similarity	PIR-ASDB: SYECSA
PIR Feature & Post Translational Modifications	FEAT1; RESID:AA0035(1'-phospho-L-histidine); RESID:AA0036(3'-phospho-L-histidine); active site:His (phosphohistidine intermediate)(247) [experimental]
PIR Alignments	PIR-ALN: FA0554: succinate–CoA ligase (ADP-forming) alpha chain
Other Classification	COG: COG0074 : sucD(1-289) BLOCKS: IPB000303: ATP-citrate lyase/succinyl-CoA ligases Pfam: PF00549: CoA-ligase(133-280) Pfam: PF02629: CoA binding domain(4-115) MetaFam: SYECSA InterPro: SUCD_ECOLI

FEATURE & SEQUENCE DISPLAY

REFRESH
PDOC00335,ATP-citrate lyase / succinyl-CoA ligases family active site
SYECSA
1
289
247
PCM00399
PCM01216
PF00549
PF02629

```
  1  MSILIDKNTKVICQGFTGSQGTFHSEQAIAYGTKMVGGVTPGKGGTTHLGLPVFNTVREA
 61  VAATGATASVIYVPAPFCKDSILEAIDAGIKLIITITEGIPTLDMLTVKVKLDEAGVRMI
121  GPNCPGVITPGECKIGIQPGHIHKPGKVGIVSRSGTLTYEAVKQTTDYGFGQSTCVGIGG
181  DPIPGSNFIDILEMFEKDPQTEAIVMIGEIGGSAEEEAAAYIKEHVTKPVVGYIAGVTAP
241  KGKRMGHAGAIIAGGKGTADEKFAALEAAGVKTVRSLADIGEALKTVLK
```

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



GENERAL INFORMATION

Superfamily Number	SF001553
Superfamily Name	succinate--CoA ligase (ADP-forming) alpha chain
Superfamily Size	Total Families=4; Total Sequence Entries=51 (49 Proteins+2 Fragments)
Taxonomy Range	Eukaryotae=9; Prokaryotae=35; Archaea=7; Viruses=0; Other=0
Length Range	Minimum=260; Maximum=368; Average=300; Standard Deviation=19

Keyword	acid-thiol ligase(22); coenzyme A(18); phosphohistidine(14); phosphoprotein(14); ATP(13);
	tricarboxylic acid cycle(10);

MEMBERSHIP

Eukaryotic Member	iProClass: S30579; S61696

Prokaryotic Member	iProClass: A70440; A85574
	S54848; S75921; SYECSA
	D64189; D72013; D82545;
	F81399; F97774; G70462;

Archaebacterial Member	iProClass: A69523; A72707
Model Organism	Caenorhabditis elegans T189 Arabidopsis thaliana T51816 Saccharomyces cerevisiae S Methanococcus jannaschii E Escherichia coli AB5574; SY

FAMILY RELATIONSHIP

PIR/MIPS Family	PIR-MIPS: FAM0001543(
PIR MOTIF	iProClass: PCM00399: ATP iProClass: PCM01216: ATP
PIR Alignment	PIR-ALN: FA0554: succina
PFAM	PFAM: PF02629: CoA bind PFAM: PF00549: CoA-liga
COG	COG: COG0074
MetaFam	MetaFam: SF001553

FUNCTION AND STRUCTURE

Enzyme	EC-IUBMB: 6.2.1.5 6.2.1.4 6.2.1.- KEGG: 6.2.1.5 6.2.1.4 BRENDA: 6.2.1.5 6.2.1.4 WIT: 6.2.1.5 6.2.1.4 MetaCyc: 6.2.1.5 6.2.1.4 6.2.1.- EcoCyc: 6.2.1.5 6.2.1.4
Structure	PDB: 1CQI; 1CQJ; 1SCU PDBSum: 1CQI;1CQJ;1SCU SCOP: 1CQI; 1CQJ; 1SCU CATH: 1CQI; 1CQJ; 1SCU FSSP: 1CQI; 1CQJ; 1SCU MMDB: 1CQI; 1CQJ; 1SCU

DOMAIN/MOTIF DISPLAY

SF001553 SUCCINATE--COA LIGASE (ADP-FORMING) ALPHA CHAIN FAMILY MEMBERS (43 sequences)

PCM00399: PDOC00335 ATP-citrate lyase / succinyl-CoA ligases family active site
PCM01216: PDOC00335 ATP-citrate lyase / succinyl-CoA ligases family signature 1
PF00549: CoA-ligase
PF02629: CoA binding domain

HELP
refresh
hide legend
A69523
246
287
PCM00399: PDOC00335, ATP-citrate lyase / succinyl-CoA ligases family active site
205
A70440
160
305
PF00549: CoA-ligase (135-282)
200
A72707
297
B69442
291
C55230
280
D64189
293

Fig. 7. The iProClass superfamily report with family relationship information.

5. Functional Associations Beyond Sequence Homology

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.[542]

The data integration in iProClass is important in revealing protein functional associations beyond sequence homology, as illustrated in the following example. As shown in Part I of Figure 8, 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 CYSN homology (PF00009+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 (Figure 8, Part II). This example demonstrates that protein function may be revealed based on domain and/or pathway association, even without obvious sequence homology. The iProClass database design presents such complex superfamily-domain-function relationships to assist functional identification or characterization of proteins.

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 grant P41 LM05978 from the National Library of Medicine, National Institutes of Health. The iProClass database is supported by DBI-9974855 from the National Science Foundation.

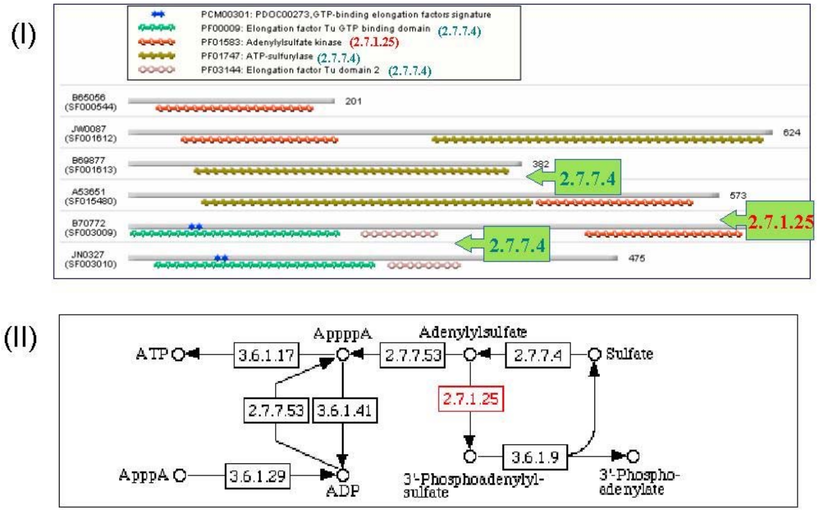

(I)
PCM00301: PDOC00273,GTP-binding elongation factors signature
PF00009: Elongation factor Tu GTP binding domain (2.7.7.4)
PF01583: Adenylylsulfate kinase (2.7.1.25)
PF01747: ATP-sulfurylase (2.7.7.4)
PF03144: Elongation factor Tu domain 2 (2.7.7.4)
B65056 (SF000544)
201
JW0087 (SF001612)
624
B69877 (SF001613)
382
2.7.7.4
A53651 (SF015480)
573
B70772 (SF003009)
2.7.1.25
JN0327 (SF003010)
2.7.7.4
475
(II)
ATP
3.6.1.17
ApppA
2.7.7.53
Adenylylsulfate
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. 8. (I) Superfamily-domain-function relationship for functional inference beyond sequence homology. Association of EC 2.7.1.25 and two distinct sequence types of EC 2.7.7.4 in multi-domain proteins. (II) Association of EC 2.7.1.25 and EC 2.7.7.4 in the same metabolic pathway.