Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry 28 (2004) 87–96
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Database note

The iProClass integrated database for protein functional analysis

Cathy H. Wu [a,*], Hongzhan Huang [a], Anastasia Nikolskaya [a], Zhangzhi Hu [b], Winona C. Barker [b]

[a] *Georgetown University Medical Center, 3900 Reservoir Road, NW, Box 571455, Preclinical Building, Room LR-3, Washington, DC 20057-1455, USA*
[b] *National Biomedical Research Foundation, 3900 Reservoir Road, NW, Washington, DC 20057-1455, USA*

Received 19 October 2003; received in revised form 21 October 2003; accepted 22 October 2003

Abstract

Increasingly, scientists have begun to tackle gene functions and other complex regulatory processes by studying organisms at the global scales for various levels of biological organization, ranging from genomes to metabolomes and physiomes. Meanwhile, new bioinformatics methods have been developed for inferring protein function using associative analysis of functional properties to complement the traditional sequence homology-based methods. To fully exploit the value of the high-throughput system biology data and to facilitate protein functional studies requires bioinformatics infrastructures that support both data integration and associative analysis. The iProClass database, designed to serve as a framework for data integration in a distributed networking environment, provides comprehensive descriptions of all proteins, with rich links to over 50 databases of protein family, function, pathway, interaction, modification, structure, genome, ontology, literature, and taxonomy. In particular, the database is organized with PIRSF family classification and maps to other family, function, and structure classification schemes. Coupled with the underlying taxonomic information for complete genomes, the iProClass system (http://pir.georgetown.edu/iproclass/) supports associative studies of protein family, domain, function, and structure. A case study of the phosphoglycerate mutases illustrates a systematic approach for protein family and phylogenetic analysis. Such studies may serve as a basis for further analysis of protein functional evolution, and its relationship to the co-evolution of metabolic pathways, cellular networks, and organisms.

Keywords: Protein functional analysis; Human genome; Bioinformatics

1. Introduction

The completion of the human genome sequences marked the beginning of a new era of biological research with rapid advances in "system biology" studies. Scientists have begun to systematically tackle gene functions and other complex regulatory processes by studying organisms at the global scales of genomes (genes, regulatory and noncoding sequences), transcriptomes (gene expression) (Carninci et al., 2003), proteomes (protein expression) (Babnigg and Giometti, 2003), metabolomes (metabolic networks) (Bono et al., 2003), interactomes (protein–protein interactions) (Walhout et al., 2002), and physiomes (physiological dynamics of whole organisms) (Hunter and Borg, 2003). Associated with the enormous quantity and variety of data being produced is the growing number of molecular databases that are being generated and maintained. Meta databases (database of databases) have been compiled to catalog and categorize these databases, such as the Molecular Biology Database Collection (Baxevanis, 2003).

To fully explore these valuable data, advanced bioinformatics infrastructures must be developed for biological knowledge management. One major challenge lies in the volume, complexity, and dynamic nature of data being collected and maintained in heterogeneous and distributed sources. To facilitate scientific discovery, information scattered in disparate sources needs to be integrated into a cohesive framework. With *data integration*, interesting relationships among protein family, structure, and function can be readily revealed, providing for plausible function and pathway identification. Indeed, protein function can be inferred using *associative analysis* ("guilt-by-association") based on system biology properties even when there is no detectable sequence similarity (Marcotte et al., 1999a; Koonin and Galperin, 2003; Osterman and Overbeek, 2003). Associative properties that have been demonstrated to allow

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

1476-9271/$ – see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2003.10.003

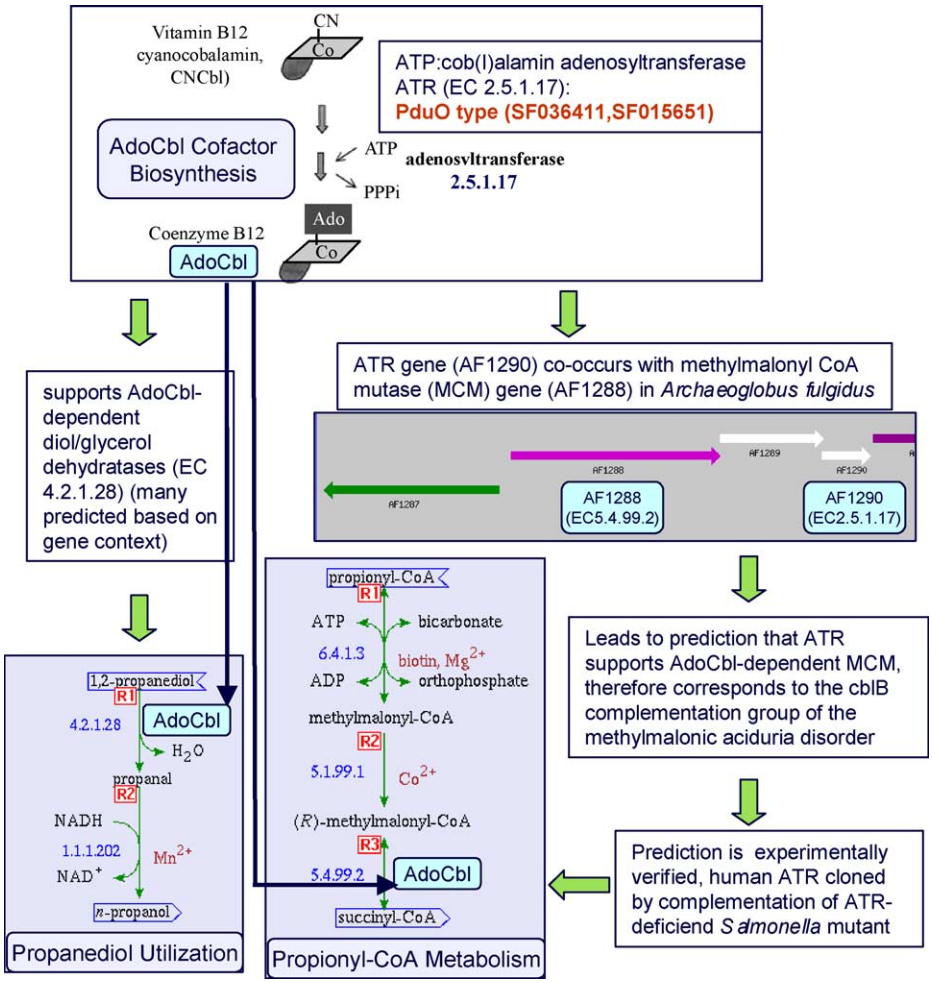

Fig. 1. Integration of protein family, pathway, and genome context data for gene identification.

inference of function not evident from sequence homology include: co-occurrence of proteins in operons or genome context (Overbeek et al., 1999); proteins sharing common domains in fusion proteins (Marcotte et al., 1999b); proteins in the same pathway, sub-cellular network, or complex; proteins with correlated gene or protein expression patterns; and protein families with correlated taxonomic distribution (common phylogenetic/phyletic patterns) (Pellegrini et al., 1999; Morett et al., 2003).

The following example (Fig. 1) shows that the collective use of protein family, pathway (enzyme and cofactor), and genome context information in bacterial organisms helps researchers identify the human gene for ATP:cob(I)alamin adenosyltransferase (ATR) (EC 2.5.1.17). ATR converts inactive cobalamins to AdoCbl (Fig. 1A), a cofactor for enzymes in several pathways, including diol/glycerol dehydratase (EC 4.2.1.28) (Fig. 1B), and methylmalonyl CoA mutase (MCM) (EC 5.4.99.2) (Fig. 1C). It has long been known that deficiencies of ATR are associated with the methylmalonic aciduria disorder (Fenton and Rosenberg, 1981), a metabolic disorder resulting from deficient MCM activity, but the ATR gene was not found. Many prokaryotic members of the ATR protein families (SF036411 and SF015651) are predicted to be required for AdoCbl-dependent diol or glycerol dehydratases, based on the genome context of the corresponding genes (Johnson et al., 2001). However, in at least one organism (*Archaeoglobus fulgidus*), the ATR gene is adjacent to sequences encoding an AdoCbl-dependent MCM homolog, which provided a clue for cloning the human and bovine ATRs (SF015651) based on sequence homology (Dobson et al., 2002).

The example indicates a clear benefit for a bioinformatics approach that supports global-scale data integration and associative analysis. This paper describes the iProClass integrated database (Huang et al., 2003) designed to provide such a framework to facilitate protein functional analysis with a case study.

2. iProClass integrated database system

2.1. Overview—rich links for data integration

The iProClass database was designed to offer a comprehensive, integrated view of protein information to facilitate

knowledge discovery and to serve as a framework for data integration in a distributed networking environment (Wu et al., 2001). There are several general approaches for developing an integrated platform for heterogeneous databases (Davidson et al., 1995). These include: hypertext navigation using links between related data sources, a data warehouse that provides a materialized solution, unmediated multi-database queries that provide view solution, and database federation. Most data warehouses adopt a "tightly coupled" approach that physically integrates a number of databases by converting the data into a unified database schema. While it allows local control of data, updating data from the multiple databases is not trivial. Hypertext navigation is a "loosely coupled" approach that employs the browsing model wherein hypertext-linked web pages are followed for more information and are always one mouse click away.

iProClass uses database links as a foundation for interoperability (Karp, 1995) and combines both data warehouse and hypertext navigation methods. In our approach, we restrict the database content to the immediate needs of protein analysis and annotation and store a rich collection of links with related summary information to alleviate potential problems associated with timely collection of information from distributed sources over the Internet. The idea is similar to that of the Virgil database (Achard et al., 1998), which was developed to model the concept of rich links (the link itself and the related summary information) between database objects. Following the notation in LinkDB (Fujibuchi et al., 1998), the iProClass links may be roughly categorized into three types: (i) factual links for simple cross-references, such as literature data or reported sequence data; (ii) similarity links compiled based on sequence similarity, such as members of a protein family; and (iii) biological links associating biological meanings, such as interacting proteins or proteins in the same metabolic pathway. Another iProClass design principle that promotes database interoperation is the adoption of a modular and open architecture. The modular structure makes the system scalable, customizable, and extendable for adding new components.

2.2. iProClass content—integrated protein information

The database contains comprehensive descriptions of all proteins with up-to-date information from many sources, thereby providing much richer annotation than can be found in any single database. The information includes protein family relationships at both whole protein and domain, motif, site levels, as well as structural and functional classifications and features of proteins. iProClass currently consists of about 1.1 million UniProt (Apweiler et al., 2004) sequence entries organized with 36,000 PIRSF families. The PIRSF (SuperFamily) classification system (Wu et al., 2004), which provides classification of whole proteins into a network structure to reflect their evolutionary relationships, is central to the iProClass database organization and the PIR/UniProt functional annotation of proteins (Wu et al., 2003a). The system is extended from the PIR superfamily/family concept (Dayhoff, 1976; Barker et al., 1996), the original classification based on sequence similarity where protein family members are homologous (sharing common ancestry) and homeomorphic (sharing full-length sequence similarity with common domain architecture).

Rich links to over 50 biological databases are provided with source attribution, hypertext links, and related summary information extracted from the underlying sources, including the following databases (please refer to the annual January issue of the *Nucleic Acids Research* for citations and updated information on these databases).

- Protein sequence: UniProt (PIR-PSD, Swiss-Prot and TrEMBL), GenPept (GenBank translations), RefSeq, PIR-NREF
- Protein families: InterPro, COG, Pfam, ProSite, Blocks, Prints, CDD, MetaFam, ProtFam
- Functions and pathways: EC-IUBMB, KEGG, BRENDA, WIT, MetaCyc, EcoCyc
- Structures and structural classifications: PDB, SCOP, CATH, MMDB, PDBsum, FSSP
- Protein-protein interactions: DIP, BIND
- Post-translational modifications: RESID, Phosphorylation Site DB
- Genes and genomes: GenBank/EMBL/DDBJ, TIGR, SGD, Flybase, MGI, GDB, OMIM, MIPS, GenProtEC, LocusLink
- Ontologies: Gene Ontology
- Literature: PubMed
- Taxonomy: NCBI Taxonomy

The source attribution and hypertext links facilitate exploration of additional information and examination of discrepant annotations from different sources. The link mechanism is also attributed. For example, EC number is used to cross-reference functional databases, PDB (Westbrook et al., 2003) ID is used to link structure and structural classification databases, and PubMed ID links to NCBI's PubMed. Standard nomenclatures and accepted ontologies are adopted wherever applicable, such as IUBMB Enzyme Nomenclature, NCBI taxonomy, and Gene Ontology (Gene Ontology Consortium, 2001).

2.3. iProClass views—value-added protein and family reports

iProClass provides value-added views for UniProt protein entries and PIRSF family entries with extensive annotation information and graphical displays. The protein summary report (Fig. 2) contains information on:

- General information: protein ID and name (with synonyms, alternative names), source organism taxonomy (with NCBI taxonomy ID, group, and lineage), and

Summary Report for iProClass Entry: PMHUYB+PMG1_HUMAN *Related Sequences*

GENERAL INFORMATION

Protein Name and ID	PIR-NREF: NF00116454

Database	ID	Accession	Protein Name
PIR-PSD	PMHUYB	A31782; A31783	phosphoglycerate mutase (EC 5.4.2.1) B [validated] *ALT_NAMES:* phosphoglycerate mutase 1; phosphoglycerate mutase brain isozyme; phosphoglycerate phosphomutase; phosphoglyceromutase
SwissProt	PMG1_HUMAN	P18669; Q9BWC0	Phosphoglycerate mutase 1 (EC 5.4.2.1) (EC 5.4.2.4) (EC 3.1.3.13) (Phosphoglycerate mutase isozyme B) (PGAM-B) (BPG-dependent PGAM 1)
RefSeq	g4505753	g4505753	phosphoglycerate mutase 1 (brain), Phosphoglycerate mutase A, nonmuscle form

GenPept: AAG01990.1 ; AAH11678.1 ; AAH10038.1 ; AAA60071.1

Taxonomy	*Source Organism:* Homo sapiens(human) *Taxon Group:* Euk/Animal NCBI Taxon: 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/Pan/Gorilla group; Homo
Gene Name	GDB:PGAM1; GDB:PGAMA
Keywords	3D-structure; acetylation; dimer; gluconeogenesis; glycolysis; hydrolase; intramolecular transferase; isomerase; phosphohistidine; phosphoprotein; phosphoric monoester hydrolase
Function	Interconversion of 3- and 2-phosphoglycerate with 2,3-bisphosphoglycerate as the primer of the reaction. Can also Catalyze the reaction of EC 5.4.2.4 (synthase) and EC 3.1.3.13 (phosphatase), but with a reduced activity
Subunit	Homodimer
Tissue Specificity	In mammalian tissues there are two types of phosphoglycerate mutase isozymes: type-M in muscles and type-B in other tissues

CROSS-REFERENCES

Bibliography	►View Bibliography Information ►Submit Bibliography PubMed: PMID: 2846553; 2846554; 6282177; 9150946; 12477932
DNA Sequence	GenBank: J04173 EMBL: J04173 DDBJ: J04173
Genome/Gene	GDB: 120530 OMIM: 172250 LocusLink: 5223 phosphoglycerate mutase 1 (brain)(PGAM1)
Ontology	*Molecular Function* GO:0003824: enzyme activity [INTERPRO; evidence:IEA] GO:0004619: phosphoglycerate mutase activity [PMID:2846554; evidence:NAS] GO:0016868: intramolecular transferase activity, phosphotransferases [INTERPRO; evidence:IEA] GO:0016787: hydrolase activity [SPKW; evidence:IEA] GO:0016853: isomerase activity [SPKW; evidence:IEA] GO:0004083: bisphosphoglycerate phosphatase activity [SPEC; evidence:IEA] *Biological Process* GO:0008152: metabolism [INTERPRO; evidence:IEA] GO:0006096: glycolysis [INTERPRO; evidence:IEA] [SPKW; evidence:IEA]
Enzyme/Function	EC 5.4.2.1 EC-IUBMB, KEGG, BRENDA, WIT, MetaCyc *Nomenclature:* Isomerases; Intramolecular Transferases; Phosphotransferases (Phosphomutases); phosphoglycerate mutase *Reaction:* 2-phospho-D-glycerate = 3-phospho-D-glycerate EC 3.1.3.13 EC-IUBMB, KEGG, BRENDA, WIT, MetaCyc *Nomenclature:* Hydrolases; Acting on Ester Bonds; Phosphoric Monoester Hydrolases; bisphosphoglycerate phosphatase *Reaction:* 2,3-bisphospho-D-glycerate + H_2O = 3-phospho-D-glycerate + phosphate EC 5.4.2.4 EC-IUBMB, KEGG, BRENDA, WIT, MetaCyc *Nomenclature:* Isomerases; Intramolecular Transferases; Phosphotransferases (Phosphomutases); bisphosphoglycerate mutase *Reaction:* 3-phospho-D-glyceroyl phosphate = 2,3-bisphospho-D-glycerate
Pathway	KEGG: Metabolism; Carbohydrate Metabolism; Glycolysis / Gluconeogenesis [PATH:hsa00010].
Structure	PDB: 1E58:A(1-253,58.5%), 1E59:A(1-253,58.5%), 4PGM:A(5-251,51.0%), 4PGM:B(5-251,51.0%), 4PGM:C(5-251,51.0%), 4PGM:D(5-251,51.0%), 1BQ3:A(5-251,51.0%), 1BQ3:B(5-251,51.0%), 1BQ3:C(5-251,51.0%), 1BQ3:D(5-251,51.0%), 1BQ4:A(5-251,51.0%), 1BQ4:B(5-251,51.0%), 1BQ4:C(5-251,51.0%), 1BQ4:D(5-251,51.0%), 5PGM:A(5-251,51.0%), 5PGM:B(5-251,51.0%), 5PGM:C(5-251,51.0%), 5PGM:D(5-251,51.0%), 5PGM:E(5-251,51.0%), 5PGM:F(5-251,51.0%), More 1BQ3: SCOP CATH FSSP MMDB PDBsum 1BQ4: SCOP CATH FSSP MMDB PDBsum 1E58: SCOP CATH FSSP MMDB PDBsum 1E59: SCOP CATH FSSP MMDB PDBsum 4PGM: SCOP CATH FSSP MMDB PDBsum 5PGM: SCOP CATH FSSP MMDB PDBsum More
PIR Feature & Post Translational Modifications	FEAT1; RESID: AA0035 (1'-phospho-L-histidine):RESID: AA0036 (3'-phospho-L-histidine)active site: His (phosphohistidine intermediate) (11) [predicted] FEAT2; active site: Arg, Arg, His (10,62,186) [predicted] FEAT3; product: phosphoglycerate mutase (2-254) [experimental]

FAMILY CLASSIFICATION

PIR FASTA Similarity	PIR-ASDB: PMHUYB
PIR Superfamily	iProClass:SF001490 phosphoglycerate mutase
PIR Family	PIR-MIPS: FAM0001464 PIR-ALN: FA0539: phosphoglycerate mutase
PIR Homology Domain	iProClass: HD00280: phosphoglycerate mutase homology(6-221) PIR-ALN: DA3761: phosphoglycerate mutase homology (6-221)
PIR Motif	iProClass: PCM00175: PDOC00158: Phosphoglycerate mutase family phosphohistidine signature (PST:8-17)
InterPro	*InterPro:* PMG1_HUMAN IPR001345: Phosphoglycerate/bisphosphoglycerate mutase IPR005952: Phosphoglycerate mutase 1
Other Classification	Pfam: PF00300: Phosphoglycerate mutase family (3-229) MetaFam: PMHUYB

FEATURE & SEQUENCE DISPLAY

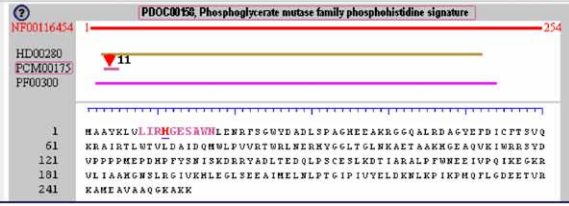

Fig. 2. iProClass protein sequence report. (This report, for a human phosphoglycerate mutase, can be viewed directly at http://pir.georgetown.edu/cgi-bin/ipcEntry?id=PMHUYB).

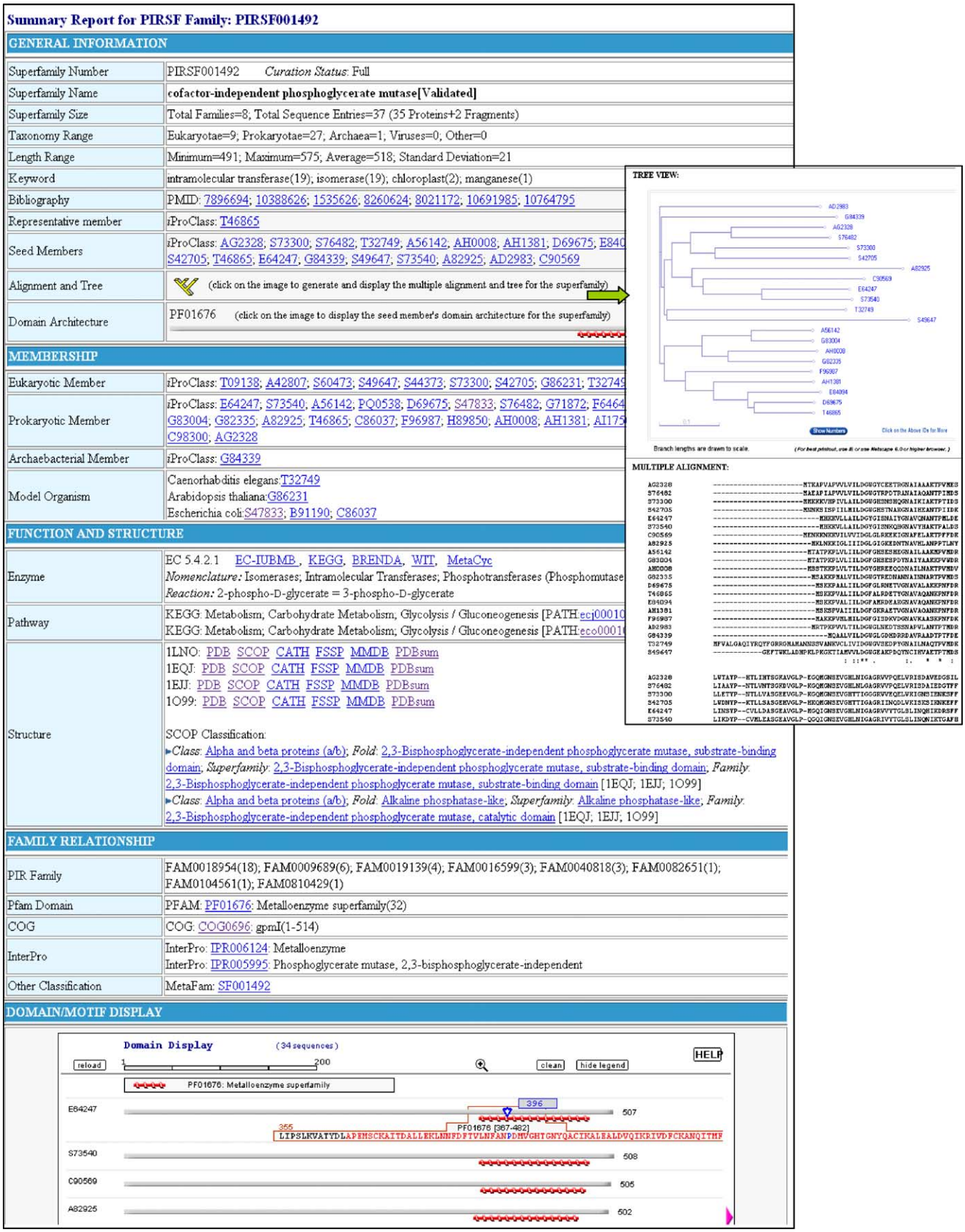

Fig. 3. iProClass protein family report. (This report, for the cofactor-dependent phosphoglycerate mutase PIRSF family, can be viewed directly at http://pir.georgetown.edu/cgi-bin/ipcSF?id=SF001492).

sequence annotations (such as gene names, keywords, function, and complex);

- Database cross-references: bibliography (with PubMed ID and link to a bibliography information and submission page), gene and genome databases, gene ontology (with GO hierarchy and evidence tag), enzyme/function (with EC hierarchy, nomenclature and reaction), pathway (with KEGG (Kanehisa et al., 2002) pathway name and link to pathway map), protein-protein interaction, structure (with PDB 3D structure image, matched residue range, and percentage sequence identity for all structures matched at ≥30% identity), structural classes (with SCOP hierarchy (Lo Conte et al., 2002) for structures at ≥90% identity), sequence features and post-translational modifications (with residues or residue ranges);
- Family classification: PIRSF family, InterPro family (Mulder et al., 2003), Pfam (Bateman et al., 2002) domain (with residue range), Prosite motif (with residue range), COG, and other classifications; and
- Sequence display: graphical display of domains and motifs on the amino acid sequence.

Family summary reports (Fig. 3) are available for PIRSF families, containing information derived from iProClass protein entries (such as membership statistics, family and function/structure relationships, and database cross-references), as well as curated family information, as summarized below:

- General information: PIRSF number and general statistics (family size, taxonomy range, length range, keywords) for preliminary clusters; additional information on family name, bibliography, family description, representative member, seed members, domain architecture, and link to multiple sequence alignment and phylogenetic tree (dynamically generated based on seed members) for curated families;
- Membership: lists of all members separated by major kingdoms and members of model organisms;
- Function, structure, and family relationship: enzyme (EC) and structure (SCOP) hierarchies, family relationships at the whole protein, domain, and motif levels with direct mapping and links to other family, function, and structure classification schemes; and
- Graphical display: domain and motif architecture of seed members or all members.

2.4. iProClass distribution—Web access and FTP download

The iProClass database is implemented in Oracle 9i database management system and updated biweekly. It is freely accessible from the PIR web site (http://pir.georgetown.edu/iproclass/) for direct report retrieval and sequence and text searches. Protein reports can be directly retrieved based on UniProt (PIR, Swiss-Prot, TrEMBL) protein sequence accession numbers or IDs (as in the example: http://pir.georgetown.edu/cgi-bin/ipcEntry?id=KMECPW). Family reports are retrievable based on PIRSF unique identifiers (as in http://pir.georgetown.edu/cgi-bin/ipcSF?id=SF001500). In addition to direct report retrieval, the protein entries are searchable by either sequence (BLAST search (Altschul et al., 1997) and peptide match) or annotation text (unique identifiers or text strings), and family entries are searchable by text. The searches return protein or family entries listed in summary lines with major annotation fields and hypertext links to full reports. Searchable text fields for both sequence and family entries include database unique identifiers (e.g., Pfam ID, EC number, and PDB ID) and annotations (e.g., family name, keywords, and sequence length). The protein report is also directly downloadable from the PIR FTP site (ftp://ftp.pir.georgetown.edu/pir_databases/iproclass/) in XML format with an associated DTD file.

3. iProClass for protein functional analysis: a case study

3.1. iProClass for integrative analysis of protein function

The data integration in iProClass facilitates functional exploration and comparative analysis of proteins. In particular, when coupled with the PIRSF classification, the iProClass system supports associative studies of protein family, domain, function, and structure. Such integrative approach with associative analysis using information on protein sequence, structure, function, and other system biology information is being employed for the protein family curation and annotation at the PIR (Wu et al., 2003b). This includes drawing on various types of available information to provide a comprehensive picture that can lead to novel predictions that can be used to show the function of the previously uncharacterized group. For example, Pfam-based searches can identify all PIRSFs sharing one or more Pfam domains. Likewise, SCOP structural classification-based searches can identify PIRSFs in the same SCOP superfamily class. Functional convergence (unrelated proteins with the same activity via non-orthologous gene displacement) and functional divergence (paralogous proteins with differing activities or expression) can be revealed by the many-to-one and one-to-many relationships between the enzyme classification (EC number) and PIRSF classification.

With the underlying taxonomic information, one can derive phylogenetic patterns of PIRSFs, indicating the presence or absence of corresponding proteins in completely sequenced genomes to identify PIRSFs that occur only in given lineages or share common phylogenetic patterns. Combining phylogenetic pattern and biochemical pathway information for protein families allows us to identify cases where alternative pathways exist for the same end product in different taxonomic groups. A well-studied example is the non-mevalonate pathway of isoprenoid biosynthesis (Fig. 4) used in pathogenic *Plasmodium falciparum*

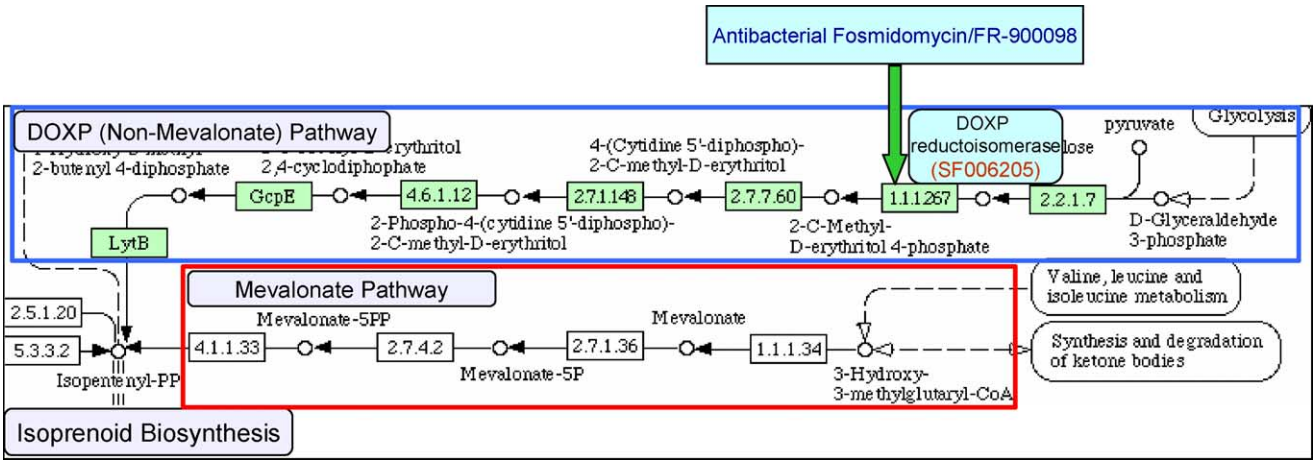

Fig. 4. Alternative pathway as a drug target.

Table 1
PIRSF family, structural classification, enzyme classification, and domain relationships

PIRSF ID	PIRSF name	Average length	SCOP superamily	Enzyme classification	pfam Domain
SF001490	Cofactor-dependent PGM	241	Phosphoglycerate mutase-like	EC 5.4.2.1; EC 5.4.2.4; EC 3.1.3.13; EC 3.1.3.-	PF00300
SF001492	Cofactor-independent PGM	518	Alkaline phosphatase-like; 2,3-BPG-independent PGM, substrate-binding domain	EC 5.4.2.1	PF01676
SF006392	Cofactor-independent PGM, archaeal type	411		EC 5.4.2.1	PF01676
SF001491	Phosphopentomutase	403		EC 5.4.2.7	PF01676
SF000891	Alkaline phosphatase	507	Alkaline phosphatase-like	EC 3.1.3.1	PF00245
SF000971	Sulfatase	580	Alkaline phosphatase-like	EC 3.1.6.8; EC 3.1.6.1; EC 3.1.6.2; EC 3.1.6.12	PF00884

PGM: phosphoglycerate mutase; BPG: bisphosphoglycerate.

and *Yersinia pestis*. It presents attractive potential drug targets because the enzymes are present in bacteria but missing in animals. Indeed, the enzyme DOXP reductoisomerase (EC 1.1.1.267) is the target of an antibacterial drug (Jomaa et al., 1999). Using integrative curation, novel cases similar to this could be identified systematically. The case study below shows how iProClass and PIRSF information can be used collectively to reveal functional convergence and divergence and analyze phylogenetic profiles.

Table 2
A PIRSF family (SF001490) with divergent functions

Protein ID	Protein name	Length	Organism
PMGE_HUMAN	Bisphosphoglycerate mutase (EC 5.4.2.4) (2,3-bisphosphoglycerate mutase, erythrocyte) (2,3-bisphosphoglycerate synthase) (BPGM) (EC 5.4.2.1) (EC 3.1.3.13) (BPG-dependent PGAM)	259	*Homo sapiens*
PMG1_HUMAN	Phosphoglycerate mutase 1 (EC 5.4.2.1) (EC 5.4.2.4) (EC 3.1.3.13) (Phosphoglycerate mutase isozyme B) (PGAM-B) (BPG-dependent PGAM 1)	254	*Homo sapiens*
PMG2_HUMAN	Phosphoglycerate mutase 2 (EC 5.4.2.1) (EC 5.4.2.4) (EC 3.1.3.13) (Phosphoglycerate mutase isozyme M) (PGAM-M) (BPG-dependent PGAM 2) (Muscle-specific phosphoglycerate mutase)	253	*Homo sapiens*
GPMB_ECOLI	Probable phosphoglycerate mutase gpmB (EC 5.4.2.1) (Phosphoglyceromutase) (PGAM)	215	*Escherichia coli*
COBC_ECOLI	Alpha-ribazole-5′-phosphate phosphatase (EC 3.1.3.-)	203	*Escherichia coli*
GPMA_ECOLI	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase (EC 5.4.2.1) (Phosphoglyceromutase) (PGAM) (BPG-dependent PGAM) (dPGM)	250	*Escherichia coli*

3.2. Functional evolution of phosphoglycerate mutases

Phosphoglycerate mutases (PGMs) are ubiquitous enzymes involved in glycolysis and gluconeogenesis. They illustrate several interesting evolutionary and biological phenomena, including evolutionary convergence and divergence, as well as multiple catalytic activities within the same molecule, and an ancient structural fold found in proteins with little sequence similarity.

3.2.1. Functional convergence

Phosphoglycerate mutase catalyzes the interconversion of 2-phosphoglycerate and 3-phosphoglycerate and is assigned the classification, EC 5.4.2.1, by the Enzyme Commission. The EC number matches three PIRSF families, SF001490, SF001492, and SF006392 (Table 1), which correspond to two unrelated forms of PGMs with differing structures and catalytic mechanisms. In the literature, they are referred to as cofactor-dependent (dPGM) and cofactor-independent

Table 3
Phylogenetic patterns of three phosphoglycerate mutases families

Kingdom	Taxonomy group	SF001490 (dPGM)	SF001492 (iPGM)	SF006392 (iPGM)
Archaea	Crenarchaeota (Desulfurococcales; Sulfolobales; Thermoproteales)	−	−	+
	Euryarchaeota (Archaeoglobi; Methanobacteria; Methanococci; Methanopyri; Thermococci; Thermoplasmata)	−	−	+
	Euryarchaeota (Halobacteria)	−	+	−
	Euryarchaeota (Euryarchaeota orders incerrae sedis)	+	+	+
Bacteria	Actinobacteria (Actinomycetales; Bifidobacteriales)	+	−	−
	Aquificae	+	−	+
	Bacteroidetes/Chlorobi group (Bacteroidetes)	+	+	+
	Bacteroidetes/Chlorobi group (Chlorobi)	+	−	−
	Chlamydiae/Verrucomicrobia group (Chlamydia; Chlamydophila)	+	−	−
	Cyanobacteria (Chroococcales)	−	+	−
	Cyanobacteria (Nostocales)	+	+	−
	Deinoccus	+	−	+
	Firmicutes (Bacilli; Clostridia)	+	+	−
	Firmicutes (Mollicutes)	−	+	−
	Fusobacteria	+	−	−
	Proteobacteria/Alphaproteobacteria (Caulobacterales)	+	−	−
	Proteobacteria/Alphaproteobacteria (Rhizobiales)	+	+	−
	Proteobacteria/Alphaproteobacteria (Rickettsiales)	−	−	−
	Proteobacteria/Betaproteobacteria (Burkholderiales; Neisseriales; Nitrosomonadales)	+	−	−
	Proteobacteria/Gammaproteobacteria (Alteromonadales; Legionellales; Vibrionales)	−	+	−
	Proteobacteria/Gammaproteobacteria (Enterobacteriales; Pseudomonadales)	+	+	−
	Proteobacteria/Gammaproteobacteria (Pateurellales; Xanthomonadales)	+	−	−
	Proteobacteria/Delta-epsilon subdivision	−	+	−
	Spirochaetes (Leptospiraceae)	−	+	−
	Spirochaetes (Spirochaetaceae)	+	−	−
	Thermotogae	+	−	+
Eukaryota	Metazoa (*Homo sapiens*; *Mus musculus*; *Rattus norvegicus*; *Danio rerio*; *Drosophila melanogaster*)	+	−	−
	Metazoa (*Caenorhabditis elegans*)	−	+	−
	Viridiplantae (*Arabidopsis thaliana*)	+	+	+
	Viridiplantae (*Oryza sativa*)	−	+	+
	Fungi (*Saccharomyces cerevisiae*)	+	−	−
	Fungi (*Encephalitozoon cuniculi*)	−	+	−
	Mycetozoa (*Dictyostelium discoideum*)	+	−	−
	Alveolata (*Plasmodium falciparum*)	+	−	−

(iPGM), with 2,3-bisphosphoglycerate (BPG) being the cofactor. As summarized in Table 1 and detailed in iProClass family reports (not shown), the dPGMs in SF001490 are single domain proteins classified in the Pfam PF00300 domain family and in the SCOP "phosphoglycerate mutase-like" fold superfamily. The iPGMs have two distantly related sequence types, represented in SF001492 (Fig. 3) and SF006392, which share a common domain PF01676 that covers the C-terminal region of the iPGM catalytic domain. Based on protein members with known structures, SF001492 maps to two SCOP fold superfamilies "alkaline phosphate-like" and "2,3-bisphosphoglycerate-independent phosphoglycerate mutase, substrate-binding domain." Other PIRSFs sharing the PF01676 domain or belonging to the "alkaline phosphate-like" fold superfamily (Table 1) are easily identifiable based on iProClass cross-references.

3.2.2. Functional divergence

Divergent function is observed within the PIRSF family SF001490 with several enzymatic activities reflected by different EC numbers (Table 1). The family has three human proteins, all having multiple catalytic activities (Table 2 and Fig. 2). As the result of four different events at the molecular level (Bond et al., 2002), each of these proteins exhibits three overall activities involving the same active site, a histidine residue that becomes phosphorylated: interconversion of 3-phosphoglycerate (3-PGA) and 2-PGA (mutase) (EC 5.4.2.1); conversion of 1,3-bisphosphoglycerate (1,3-BPG) to 2,3-BPG (synthase) (EC 5.4.2.4); and hydrolysis of 2,3-BPG to 2- or 3-PGA and phosphate (phosphatase) (EC 3.1.3.13). The family also has three *Escherichia coli* proteins, two functioning as dPGMS, while the third is an alpha-ribazole-5′-phosphate phosphatase (EC 3.1.3.-) involved in cobalamin biosynthesis.

3.2.3. Phylogenetic pattern

The phylogenetic patterns of the three PGM families (Table 3) were compiled based on the taxonomic hierarchy of over 200 complete genomes representing 60 taxonomic groups. The patterns show that all but one group of organisms (namely the Rickettsiales, represented by *Rickettsia prowazekii* and *Rickettsia conorii*) have at least one form of PGM, indicating that EC 5.4.2.1 is an essential enzyme for all organisms having the glycolysis/gluconeogenesis metabolism. Indeed, *Rickettsia* is an obligate parasite with a reduced genome—it uses the ATP of the host, missing all genes that support anaerobic glycolysis (Andersson et al., 1998). The phylogenetic patterns also reveal that the dPGM (SF001490) is the only form used by many higher organisms (Eukaryotes), and that most organisms use the SF001492 type iPGM. The second type of iPGM (SF006392) is found in most archaea and a few bacteria (thermophilic bacteria *Aquifex* and *Thermotoga* and radio-resistant bacteria *Deinococcus*) and plants (*Arabidopsis* and rice).

4. Conclusions

The large volume and complexity of biological data being generated represents both a challenge and an opportunity for bioinformatics research and development. To maximize the utlization of these valuable data for scientific discovery, information needs to be integrated into a cohesive framework. Data integration facilitates exploration, 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 discovered. Such associative analysis of various properties of proteins provides a comprehensive picture that can lead to novel prediction and functional inference for previously uncharacterized "hypothetical" proteins and protein groups. The case study illustrates that a systematic approach to protein family and phylogenetic analysis, supported by an integrated bioinformatics framework, may serve as a basis for further analysis and exploration of protein function and evolution. Such knowledge is fundamental to system biology studies at various levels of biological organization, ranging from genes to genomes, enzymes to metabolic pathways, and organisms to communities.

Acknowledgements

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

References

Achard, F., Cussat-Blanc, C., Viara, E., Barillot, E., 1998. The new Virgil database: a service of rich links. Bioinformatics 14, 342–348.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Gasteiger, E., Huang, H., Martin, M.J., Natale, D.A., O'Donovan, C., Yeh, L.-S., 2004. UniProt: universal protein knowledgebase. Nucleic Acids Res. 32, (in press).

Babnigg, G., Giometti, C.S., 2003. ProteomeWeb: a web-based interface for the display and interrogation of proteomes. Proteomics 3, 584–600.

Barker, W.C., Pfeiffer, F., George, D.G., 1996. Superfamily classification in PIR-international protein sequence database. Methods Enzymol. 266, 59–71.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E., 2002. The Pfam protein families database. Nucleic Acids Res. 30, 276–280.

Baxevanis, A.D., 2003. The molecular biology database collection: 2003 update. Nucleic Acids Res. 31, 1–12.

Bond, C.S., White, M.F., Hunter, W.N., 2002. Mechanistic implications for *Escherichia coli* cofactor-dependent phosphoglycerate mutase based on the high-resolution crystal structure of a vanadate complex. J. Mol. Biol. 316, 1071–1081.

Bono, H., Nikaido, I., Kasukawa, T., Hayashizaki, Y., Okazaki, Y., 2003. Comprehensive analysis of the mouse metabolome based on the transcriptome. Genome Res. 13, 1345–1349.

Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Hayashizaki, Y., 2003. Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. Genome Res. 13, 1273–1289.

Davidson, S.B., Overton, C., Nuneman, P., 1995. Challenges in integrating biological data sources. J. Comp. Biol. 2, 557–572.

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

Dobson, C.M., Wai, T., Leclerc, D., Kadir, H., Narang, M., Lerner-Ellis, J.P., Hudson, T.J., Rosenblatt, D.S., Gravel, R.A., 2002. Identification of the gene responsible for the cblB complementation group of vitamin B12-dependent methylmalonic aciduria. Hum. Mol. Genet. 11, 3361–3369.

Fenton, W.A., Rosenberg, L.E., 1981. The defect in the cbl B class of human methylmalonic acidemia: deficiency of cob(I)alamin adenosyltransferase activity in extracts of cultured fibroblasts. Biochem. Biophys. Res. Commun. 98, 283–289.

Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., Kanehisa, M., 1998. DBGET/LinkDB: an integrated database retrieval system. Pac. Symp. Biocomput. 683–694.

Gene Ontology Consortium, 2001. Creating the gene ontology resource: design and implementation. Genome Res. 11, 1425–1433.

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, 390–392.

Hunter, P.J., Borg, T.K., 2003. Integration from proteins to organs: the Physiome Project. Nat. Rev. Mol. Cell Biol. 4, 237–243.

Johnson, C.L., Pechonick, E., Park, S.D., Havemann, G.D., Leal, N.A., Bobik, T.A., 2001. Functional genomic, biochemical, and genetic characterization of the Salmonella pduO gene an ATP:cob(I)alamin adenosyltransferase gene. J. Bacteriol. 183, 1577–1584.

Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K., Soldati, D., Beck, E., 1999. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 285, 1573–1576.

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

Karp, P.D., 1995. A strategy for database interoperation. J. Comp. Biol. 2, 573–586.

Koonin, E.V., Galperin, M.Y., 2003. *Sequence—Evolution—Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers, Boston, MA. pp. 461

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., 1999a. Combined algorithm for genome-wide prediction of protein function. Nature 402, 83–86.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999b. Detecting protein function and protein-protein interactions from genome sequences. Science 285, 751–753.

Morett, E., Korbel, J.O., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B., Bork, P., 2003. Systematic discovery of analogous enzymes in thiamin biosynthesis. Nat. Biotechnol. 21, 790–795.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R.R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S.E., Pagni, M., Peyruc, D., Ponting, C.P., Selengut, J.D., Servant, F., Sigrist, C.J., Vaughan, R., Zdobnov, E.M., 2003. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res. 31, 315–318.

Osterman, A., Overbeek, R., 2003. Missing genes in metabolic pathways: a comparative genomics approach. Curr. Opin. Chem. Biol. 7, 238–251.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. U.S.A. 96, 2896–2901.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O., 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U.S.A. 96, 4285–4288.

Walhout, A.J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K.C., Schetter, A.J., Morton, D.G., Kemphues, K.J., Reinke, V., Kim, S.K., Piano, F., Vidal, M., 2002. Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. Curr Biol. 12, 1952–1958.

Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H.M., 2003. The Protein Data Bank and structural genomics. Nucleic Acids Res. 31, 489–491.

Wu, C.H., Xiao, C., Hou, Z., Huang, H., Barker, W.C., 2001. iProClass: an integrated, comprehensive and annotated protein classification database. Nucleic Acids Res. 29, 52–54.

Wu, C.H., Huang, H., Yeh, L.-S., Barker, W.C., 2003a. Protein family classification and functional annotation. Comp. Biol. Chem. 27, 37–47.

Wu, C.H., Yeh, L.-S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E., Vinayaka, C.R., Zhang, J., Barker, W.C., 2003b. The Protein Information Resource. Nucleic Acids Res. 31, 345–347.

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.-.S., Natale, D., Vinayaka, C.R., Hu, Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R.S., Suzek, B.E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J.L., Chung, S., Castro-Alvear, J., Dinkov, G., Barker, W.C., (2004). PIRSF family classification system at the Protein Information Resource. Nucleic Acids Res., 32, (in press).