D438–D442 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh100

MetaCyc: a multiorganism database of metabolic pathways and enzymes

Cynthia J. Krieger, Peifen Zhang[1], Lukas A. Mueller[2], Alfred Wang, Suzanne Paley, Martha Arnaud, John Pick, Seung Y. Rhee[1] and Peter D. Karp*

SRI International, 333 Ravenswood, Menlo Park, CA 94025, USA, [1]Carnegie Institution of Washington, 260 Panama Street, Stanford, CA 94305, USA and [2]Cornell University, Emerson Hall Room 251, Ithaca, NY 14853, USA

Received September 15, 2003; Revised and Accepted October 8, 2003

ABSTRACT

The MetaCyc database (see URL http://MetaCyc.org) is a collection of metabolic pathways and enzymes from a wide variety of organisms, primarily microorganisms and plants. The goal of MetaCyc is to contain a representative sample of each experimentally elucidated pathway, and thereby to catalog the universe of metabolism. MetaCyc also describes reactions, chemical compounds and genes. Many of the pathways and enzymes in MetaCyc contain extensive information, including comments and literature citations. SRI's Pathway Tools software supports querying, visualization and curation of MetaCyc. With its wide breadth and depth of metabolic information, MetaCyc is a valuable resource for a variety of applications. MetaCyc is the reference database of pathways and enzymes that is used in conjunction with SRI's metabolic pathway prediction program to create Pathway/Genome Databases that can be augmented with curation from the scientific literature and published on the world wide web. MetaCyc also serves as a readily accessible comprehensive resource on microbial and plant pathways for genome analysis, basic research, education, metabolic engineering and systems biology. In the past 2 years the data content and the Pathway Tools software used to query, visualize and edit MetaCyc have been expanded significantly. These enhancements are described in this paper.

INTRODUCTION

An initial motivation for creating MetaCyc (1), a multiorganism database (DB) of metabolic pathways and enzymes, was to provide a reference database to be used in conjunction with SRI International's (SRI's) Pathway Tools software (2) to computationally predict the metabolic pathway complement of an organism from its annotated genome. The Pathway Tools software is then used to create a Pathway/Genome Database (PGDB) from the resulting collection of predicted pathways. MetaCyc also serves as a comprehensive resource on microbial and plant metabolism that is readily accessible via the world wide web (see URL http://MetaCyc.org). MetaCyc is a member of the BioCyc collection of PGDBs (see URL http://BioCyc.org).

MetaCyc was initialized with all the metabolic pathways in EcoCyc (3), a model organism DB (MOD) for *Escherichia coli*. Since then, metabolic pathways from a variety of organisms have been added to MetaCyc, including those in Eukarya and Archaea. Species that have four or more pathways represented in MetaCyc are listed in Table 1. To date, most of the pathways in MetaCyc are microbial and plant pathways. An example plant pathway including comments is shown in Figure 1 and can be viewed at URL http://biocyc.org:1555/META/NEW-IMAGE?type=PATHWAY&object=PWY-882.

The Pathway Tools software used for querying, visualization and curation of MetaCyc and other pathway databases has also been extensively enhanced in the past 2 years. This paper describes the recent enhancements to MetaCyc and the software, and how to access them both. Please cite this paper when referencing MetaCyc.

INCREASED AND DIVERSIFIED DATA CONTENT

The curation efforts for MetaCyc increased significantly in the past 2 years, resulting in a significant increase in the number of database objects (see Table 2). Since MetaCyc was initialized with all the metabolic pathways from EcoCyc, most of the central metabolic pathways ubiquitous to microorganisms were already well represented in MetaCyc. Thus, the curation strategy was to add new pathways that would provide breadth. To diversify MetaCyc's curation expertise, SRI began a collaboration with The *Arabidopsis* Information Resource (TAIR) at the Carnegie Institution of Washington (Carnegie), which presently curates plant pathways, while SRI continues to curate microbial pathways.

Together, SRI and Carnegie developed a curation strategy for adding new pathways and editing existing pathways. New pathways are curated in the following order: first, central metabolic pathways universal to plants; second, secondary metabolic pathways and other pathways shared among fewer

*To whom correspondence should be addressed. Tel: +1 650 859 4358; Fax: +1 650 859 3735; Email: pkarp@ai.sri.com

Table 1. List of species that have four or more experimentally elucidated pathways represented in MetaCyc

Species	Number of pathways
Bacteria	
Escherichia coli	156
Salmonella typhimurium	20
Bacillus subtilis	20
Pseudomonas putida	14
Haemophilus influenzae	13
Deinococcus radiodurans	10
Mycoplasma capricolum	9
Pseudomonas aeruginosa	9
Mycoplasma capricolum	9
Mycoplasma pneuomiae	7
Thauera aromatica	6
Mycobacterium tuberculosis	5
Thermotoga maritima	4
Klebsiella pneumoniae	4
Eukarya	
Arabidopsis thaliana	47
Homo sapiens	30
Saccharomyces cerevisiae	14
Glycine max	11
Archaea	
Sulfolobus solfataricus	18
Methanococcus jannaschii	4

The species are grouped by taxonomic domain and are ordered within each domain based on the number of MetaCyc pathways labeled with the given species. MetaCyc pathways may be labeled with a higher-level taxon, such as genus, if all the species within that genus are thought to have the given pathway. Higher-level taxa are not included in this table.

microorganisms or plants; and, third, significant pathways restricted to only a few microorganisms or plants. Existing MetaCyc pathways that contained few to no comments or lacked enzymes are reviewed, and comments, enzyme information, and literature citations are added, redundant pathways are deleted, and errors are corrected.

To ensure the consistency of curation procedures over time and among multiple curators, our curation procedures were refined and documented in our evolving Pathway Tools Curators' Guide (see URL http://bioinformatics.ai.sri.com/ptools/curatorsguide.pdf). The Pathway Tools Curators' Guide documents the type of information that should be collected and entered for each pathway, reaction, enzyme, gene and chemical compound, and describes stylistic conventions.

We have recently increased the number of MetaCyc releases from two to four times per year to distribute the new data content more quickly. Each release includes a list of the salient changes.

Pathways, enzymes, reactions and compounds

Literature curation is very time consuming. However, MetaCyc offers a unique paradigm for significantly increasing the rate at which its data content can be expanded: PGDBs that are created using MetaCyc and Pathway Tools software can be augmented with manual curation by outside groups, and then the newly curated pathways and enzymes can be imported into MetaCyc (see Fig. 2). As the database content in MetaCyc grows, the better MetaCyc will serve as a reference DB for metabolic pathway prediction, and as a comprehensive resource on metabolic information. For example, the addition of plant pathways to MetaCyc significantly increases the capability to predict plant-specific pathways.

Since spring 2002, 29 new microbial pathways and 36 new plant pathways have been added to MetaCyc. The 29 microbial pathways add breadth to the existing microbial pathways, while the 36 plant pathways nearly complete the central metabolic pathways universal to plants and include important plant secondary metabolic pathways. Many of these new pathways were curated directly in MetaCyc while others were imported into MetaCyc from other PGDBs, such as EcoCyc, MtbRvCyc, a PGDB for *Mycobacterium tuberculosis* H37RV curated by Stanford University, and AraCyc, a PGDB for *Arabidopsis thaliana* curated by TAIR (4). We are also collaborating with the *Saccharomyces* Genome Database (SGD), which has created a pathway database for *Saccharomyces cerevisiae* that is actively being curated; we plan on importing their newly curated pathways and enzymes into MetaCyc. We also hope to import new pathways from other PGDBs; ~35 groups in academia and industry have licensed SRI's Pathway Tools software and have created, or are in the process of creating PGDBs.

In addition to curating new pathways, ~20 microbial pathways and 10 plant pathways that already existed in MetaCyc were edited extensively, which included adding comments, enzymes and literature citations. The addition of enzymes improves pathway prediction because enzymes from the annotated genome are matched to pathways in MetaCyc by enzyme name or EC number (2). Comments also aid pathway prediction and improve MetaCyc as a comprehensive resource by explaining the physiological role of pathways. Duplicate pathways are deleted to minimize redundancy.

We have added ~460 enzymes to MetaCyc since version 5.6, which was released in 2001 and described in a previous *Nucleic Acids Research* paper (1). Enzyme-specific information described in MetaCyc includes cofactors, prosthetic groups, activators, inhibitors, substrate specificity, subunit composition, comments and literature citations.

MetaCyc was also updated to reflect additions and changes to the Enzyme Nomenclature (i.e. the EC system) by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) by incorporating version 30.0 of the EC system.

We have added ~600 chemical structures to MetaCyc since version 5.6. Chemical structures help users to visualize chemical transformations in pathways and permit automated consistency checking of MetaCyc reactions.

Taxonomies

MetaCyc contains several taxonomies including the EC system (5), a pathway class hierarchy and a compound class hierarchy. We recently updated and improved the pathway class hierarchy, which can be accessed at URL http://biocyc.org:1555/META/class-subs?object=Pathways. We continue to expand and enhance MetaCyc's taxonomies to facilitate browsing and to better represent the new information added to MetaCyc, such as new classes of pathways and compounds, and the new information related to higher eukaryotes, such as subcellular compartment localization.

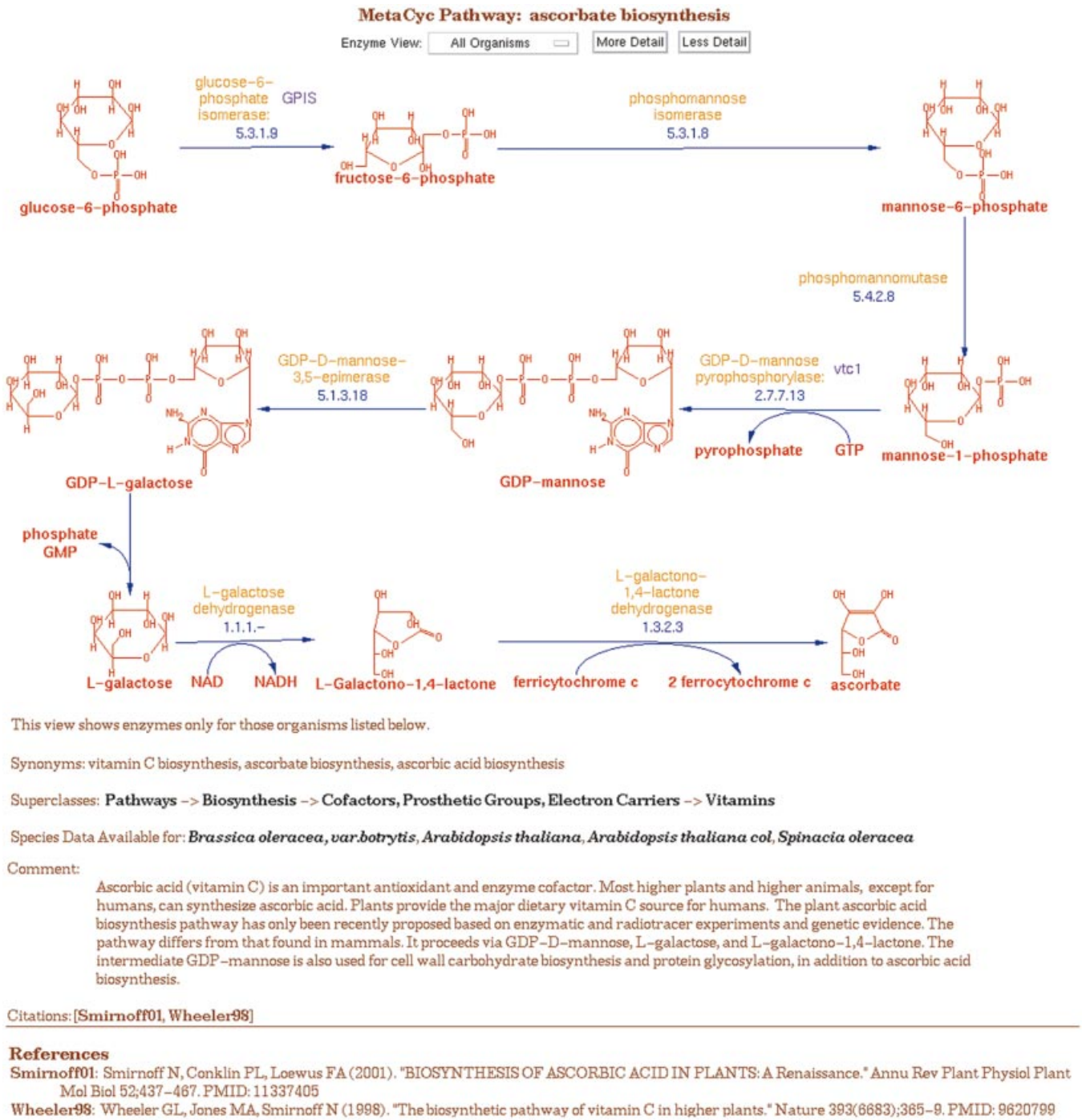

Figure 1. A representative example of a plant pathway in MetaCyc. Pathways can be displayed at various levels of detail. This pathway display depicts the greatest level of detail including enzymes, EC numbers, genes and chemical structures.

Links to other databases

MetaCyc contains unidirectional and bidirectional links to many bioinformatics databases, enabling easy navigation to and from additional biological information. MetaCyc is linked to the protein sequence databases, Swiss-Prot (6) and Protein Information Resource (PIR) (7), to the protein structural database, Protein Data Bank (PDB) (8) and to TAIR (9). We will also establish links from *S.cerevisiae* pathways in MetaCyc to SGD (10).

ENHANCEMENTS TO THE PATHWAY TOOLS SOFTWARE

The PathoLogic component of Pathway Tools (2,11) creates a new PGDB for an organism. PathoLogic populates the PGDB

Table 2. The size of MetaCyc as a function of time from its first release in 1999 to its most recent release in 2003

Database object	1999	2000	2001	2002	2003
Metabolic pathways	296	366	445	460	491
Metabolic pathways with comments	39	83	160	180	232
Enzymic reactions	3779	4002	4218	4294	4817
Enzymes	82	344	1115	1267	1543
Enzymes with comments	75	234	1054	1123	1389
Genes	0	0	0	600	1554
Compounds	1949	2180	2335	2404	2951
Literature citations	184	604	2381	2718	3070

Each row depicts the number of different database objects in MetaCyc during the last release of that year.

with the genome of the organism (obtained from a GenBank entry), the predicted metabolic pathways of the organism and predicted operons for bacteria (a recent enhancement).

The Pathway/Genome Editors component allows curators to manually refine a PGDB, such as incorporating literature-derived information. Individual editors exist for all Pathway Tools data types, including pathways, reactions, genes, enzymes, operons and chemical compounds. Recent enhancements include the import/export utility, which can be used to transfer a pathway (including the enzymes, reactions and substrates affiliated to the pathway) from one PGDB to another.

The Pathway/Genome Navigator component provides query, visualization, analysis and web-publishing operations. A recent enhancement is that the Navigator display windows (such as the pathway display) now feature full lists of references at the bottom of the page, in addition to the reference links that existed previously.

We also made several enhancements that affect more than one of the preceding software components. We implemented an Evidence Code ontology that curators can use to describe the type of evidence for the existence of an object such as a pathway or an operon. For example, an evidence code can be used to describe whether a pathway was predicted computationally or elucidated experimentally. Icons in MetaCyc display windows distinguish between computational and experimental evidence for an assertion and provide access to information regarding the evidence for a given entity. Evidence information will be available in subsequent MetaCyc releases. Pathway Tools has been extended to support editing and visualization of introns, exons, alternative splicing and annotation of protein active sites and domains. New Perl (4) and Java Application Program Interfaces (APIs) now support programmatic queries and updates to Pathway Tools PGDBs (see URL http://arabidopsis.org/tools/aracyc/javacyc for the Java API developed by Thomas Yan).

DATABASE AND SOFTWARE AVAILABILITY

MetaCyc is available free of charge via the WWW at http://Metacyc.org/ (updated four times per year). It is also available free of charge to non-profit organizations or for a fee to commercial institutions as an application program for Linux/Intel, Windows/Intel and the Sun workstation (updated twice per year), and as a set of flat files (updated four times per year).

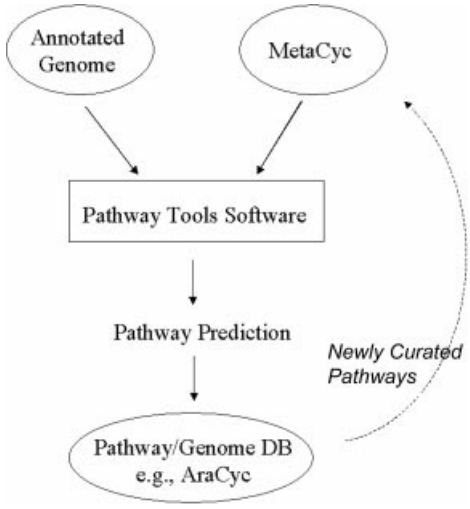

Figure 2. MetaCyc is the reference database of pathways and enzymes that is used in conjunction with SRI's Pathway Tools software to predict metabolic pathways from an organism's annotated genome, resulting in the creation of a PGDB, such as AraCyc. PGDBs can then be augmented with literature curation and the newly curated pathways and enzymes can be imported into MetaCyc, expanding its data content.

For more access information, contact metacyc-info@ai.sri.com.

ACKNOWLEDGEMENTS

We are grateful to John Ingraham for his contributions to reviewing MetaCyc pathways and for his assistance in developing the MetaCyc pathway taxonomy. We thank Dr Monica Riley for extensive contributions to the curation of MetaCyc. We thank Brett Kawakami and Tom Butler for contributions to the chemical compound data within MetaCyc, and Thomas Yan for developing JavaCyc. This work was supported by grant R01-RR07861-01 from the National Institutes of Health (NIH) National Institute for General Medical Sciences (MetaCyc), and by grant R01-HG02729-01 from the NIH National Human Genome Research Institute. L.M. and S.Y.R. were supported in part by NSF grant DBI-9978564. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. This is Carnegie publication 1628.

REFERENCES

1. Karp,P.D., Riley,M., Paley,S.M. and Pellegrini-Toole,A. (2002) The MetaCyc Database. *Nucleic Acids Res.*, **30**, 59–61.
2. Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18**, S225–232.
3. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.
4. Mueller,L.A., Zhang,P. and Rhee,S.Y. (2003) AraCyc: a biochemical pathway database for *Arabidopsis*. *Plant Physiol.*, **132**, 453–460.
5. Webb,E. (1992) *Enzyme Nomenclature 1992: Recommendations of the nomenclature committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, San Diego, CA.
6. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.*

(2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
7. Wu,C.H., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z.Z., Ledley,R.S., Lewis,K.C., Mewes,H.W., Orcutt,B.C. *et al.* (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.
8. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
9. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
10. Weng,S., Dong,Q., Balakrishnan,R., Christie,K., Costanzo,M., Dolinski,K., Dwight,S.S., Engel,S., Fisk,D.G., Hong,E. *et al.* (2003) *Saccharomyces* Genome Database (SGD) provides biochemical and structural information for budding yeast proteins. *Nucleic Acids Res.*, **31**, 216–218.
11. Paley,S.M. and Karp,P.D. (2002) Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. *Bioinformatics*, **18**, 715–724.