Nucleic Acids Research, 2002, Vol. 30, No. 1 59–61

The MetaCyc Database

Peter D. Karp*, Monica Riley[1], Suzanne M. Paley and Alida Pellegrini-Toole[1]

Bioinformatics Research Group, SRI International, 333 Ravenswood Avenue EK207, Menlo Park, CA 94025, USA and [1]Marine Biological Laboratory, Woods Hole, MA 02543, USA

Received September 18, 2001; Revised and Accepted October 17, 2001

ABSTRACT

MetaCyc is a metabolic-pathway database that describes 445 pathways and 1115 enzymes occurring in 158 organisms. MetaCyc is a review-level database in that a given entry in MetaCyc often integrates information from multiple literature sources. The pathways in MetaCyc were determined experimentally, and are labeled with the species in which they are known to occur based on literature references examined to date. MetaCyc contains extensive commentary and literature citations. Applications of MetaCyc include pathway analysis of genomes, metabolic engineering and biochemistry education. MetaCyc is queried using the Pathway Tools graphical user interface, which provides a wide variety of query operations and visualization tools. MetaCyc is available via the World Wide Web at http://ecocyc.org/ecocyc/metacyc.html, and is available for local installation as a binary program for the PC and the Sun workstation, and as a set of flatfiles. Contact metacyc-info@ai.sri.com for information on obtaining a local copy of MetaCyc.

INTRODUCTION

The MetaCyc database is an online reference source for metabolic data. It describes metabolic pathways, reactions, enzymes and substrate compounds. MetaCyc is a review-level database in that a given entry in MetaCyc often integrates information from multiple literature sources.

Intended uses of MetaCyc include the following. MetaCyc is a general reference source on metabolic pathways for the scientific community. It serves as a reference pathway database for prediction of the pathway complement of an organism from its annotated genome (1,2). MetaCyc is used as an aid in teaching biochemistry and is a resource for metabolic engineering. The modification of a metabolic network through genetic engineering involves (i) inserting a new enzyme or pathway into an organism, (ii) replacing an existing enzyme or pathway with a substitute or (iii) removing an enzyme or pathway. Cases (i) and (ii) both involve insertion of one or more new enzymes into an organism, such as enzymes with different kinetic properties, with different substrate-level regulation properties, or with new catalytic functions not present in the host. MetaCyc provides a searchable encyclopedia of enzymes and pathways that lists not only the catalytic function of enzymes, but also in many cases their substrate-level regulation (to ensure that the enzyme will be active under the appropriate cellular conditions) and cofactor requirements (so that appropriate cofactor-biosynthesis pathways can be engineered to satisfy the cofactor requirements of an inserted enzyme).

MetaCyc aims to provide a large selection of pathways from many organisms. The philosophy of MetaCyc is to encode pathways that have been reported in the experimental literature. Each pathway is labeled with the organism(s) in which it is known to occur, based on wet-lab experiments reported in the literature evaluated to date. Because experimentalists have demonstrated the presence of most pathways in only a small fraction of the organisms in which they actually occur, and because MetaCyc does not cover all known literature articles, the species information in MetaCyc is incomplete, yet it reflects wet-lab rather than computational determinations. MetaCyc employs the same database schema as does EcoCyc. It aims to provide the same rich literature-based annotation for each pathway as does EcoCyc, although a minority of pathways currently lack the extensive commentary and literature citations that we plan to provide. Each MetaCyc pathway contains a citation to the source from which it was obtained. Unlike EcoCyc, MetaCyc does not provide genomic data such as genomic maps or sequences.

This article describes recent enhancements to MetaCyc, and how to access the database. We request that users of MetaCyc cite this article in publications related to their use. Version 5.6 of MetaCyc was released in June 2001.

THE MetaCyc DATA

Table 1 shows the current sizes of the principal MetaCyc classes. The most common organisms from which MetaCyc pathways are derived are listed in Table 2. The most frequently occurring organism in MetaCyc is *Escherichia coli* because MetaCyc contains all metabolic pathways and enzymes from the EcoCyc database. The MetaCyc data were gathered from a variety of literature and online sources.

Pathways

MetaCyc does not contain redundant entries for the same metabolic pathway in the sense that no two pathways in MetaCyc contain the same set of reaction steps connected in the same topology. Separate observations of the same pathway in new organisms do expand the species distribution recorded for that pathway, but those new observations do not result in new pathway records in MetaCyc unless the pathway differs in its component reactions.

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

Table 1. The number of objects in version 5.6 of MetaCyc

Metabolic Pathways	445
Reactions	4218
Enzymes	1115
Compounds	2335
Citations	2381

MetaCyc groups together related *pathway variants* into a common class of pathways. For example, the class *Arginine Degradation* contains 13 different pathways for the degradation of arginine. In some cases variant pathways will share many reaction steps in common and will produce the same end product; in other cases the pathways will share no reaction steps in common and will produce different end products.

Enzymes

MetaCyc contains extensive information on many enzymes including descriptions of enzyme subunit structure; activators, inhibitors, cofactors and prosthetic groups; alternative substrates; explanatory comments; and citations. The species from which the enzyme information was obtained is usually the same as the pathway to which the enzyme is attached, and is recorded in the database. In some cases the database describes isozymes from different species that catalyze the same reaction.

Reactions

MetaCyc contains all reactions in the enzyme-classification system devised by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), last refreshed from version 25.0 of the ENZYME database. But, because many known reactions have not been captured by the Enzyme Commission (EC) system to date, more than 400 reactions in MetaCyc do not have an assigned EC number. It also contains thousands of objects representing individual metabolites, of which 1860 have chemical structures.

Database links

MetaCyc contains URL-based links to the PIR protein-sequence database, to the ENZYME database (which links to SWISS-PROT) and to PubMed.

PATHWAY TOOLS SOFTWARE ENVIRONMENT

The MetaCyc data reside within the software environment used for EcoCyc: the Pathway Tools (http://bioinformatics.ai.sri.com/ptools/). The Pathway Tools run in both a World Wide Web mode and an X-windows mode. All the visualization and query tools available for EcoCyc are also available for MetaCyc. For example, users can query metabolic pathways, enzymes and substrates by exact name or by substring search. Users can query pathways, substrates and reactions by taxonomies of these entities, such as the EC taxonomy of enzyme-catalyzed reactions. Query answers are displayed using software that creates graphical depictions of metabolic pathways, enzymes, reactions and substrates. For example, the pathway drawing software can display linear, circular and tree-structured pathways at multiple levels of detail.

The Pathway Tools software component called PathoLogic uses the MetaCyc database to predict the metabolic-pathway complement of an organism from its genome (2). PathoLogic assesses the evidence for the presence of each MetaCyc pathway in the organism under analysis, and creates a new pathway/genome database that models the pathways and genome of that organism. Eight such bacterial databases are available at http://ecocyc.org/.

Table 2. The organisms in which MetaCyc pathways were most frequently observed, and the number of MetaCyc pathways from those organisms

Escherichia coli	173
Salmonella typhimurium	35
Homo sapiens	31
Sulfolobus solfataricus	20
Bacillus subtilis	18
Soybean	18
Pseudomonas	17
Haemophilus influenzae	15
Mycoplasma capricolum	12
Saccharomyces cerevisiae	8
Pseudomonas putida	7
Mycoplasma pneumoniae	7
Ascomycotina	6
Rhizobiaceae	5
Clostridium	4
Pseudomonas aeruginosa	4
Thauera aromatica	4
Thermotoga maritima	4
Rhodococcus	4
Klebsiella pneumoniae	4
Archaea	3
Pseudomonadacea	3
Pseudomonas sp	3
Neisseriaceae	3
Klebsiella aerogenes	3
Rattus norvegicus	3
Archaebacteria	3
Methanosarcina barkeri	3
Sinorhizobium meliloti	3

COMPARISON OF MetaCyc AND KEGG

We compare MetaCyc to the KEGG database to provide users with a sense of the relative strengths and weaknesses of these two resources.

MetaCyc contains extensive comments that describe individual pathways and enzymes. KEGG has no comments.

MetaCyc cites the primary literature sources from which pathway and enzyme data were obtained. KEGG contains no literature citations.

MetaCyc pathways are typically smaller than KEGG pathways because KEGG typically combines together in one pathway diagram a number of related pathways from several different species (MetaCyc superpathways perform a similar function in allowing the user to view the interconnections among several pathways). MetaCyc records separately the different pathway variants that have been observed in different organisms; KEGG does not explicitly record pathway variants.

MetaCyc pathways are labeled with information regarding the species in which the presence of those pathways has been experimentally determined, whereas KEGG contains no information about which pathways or pathway fragments have been observed in which species. KEGG does allow the user to easily view, for a given pathway, which enzymatic steps in that pathway are predicted to occur in many sequenced genomes, which MetaCyc does not.

MetaCyc contains data on enzyme properties for specific enzymes from specific species, such as subunit composition, substrate specificity, cofactor requirements, activators and inhibitors. KEGG contains none of these data.

DISTRIBUTION

MetaCyc is available in four forms:

1. It is accessible online via the World Wide Web at http://ecocyc.org/ecocyc/metacyc.html (this version supports a subset of the GUI functionality of the X-windows and PC versions).
2. An X-windows version of MetaCyc for the Sun workstation bundles together the Pathway/Genome Navigator software with the MetaCyc DB.
3. A new PC version of MetaCyc bundles together the Pathway/Genome Navigator software with the MetaCyc DB.
4. A flatfile version of MetaCyc is available for global analyses.

All four forms of access are free to academic institutions for research use (contact metacyc-info@ai.sri.com for information on obtaining MetaCyc). A fee applies to commercial use. The EcoCyc/MetaCyc World Wide Web site provides background information about the databases and software, and access to the publications produced by the EcoCyc and MetaCyc projects.

ACKNOWLEDGEMENTS

This work was supported by grant 1-R01-RR07861-01 from the Comparative Medicine Program at the National Center for Research Resources. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

REFERENCES

1. Karp,P.D., Ouzounis,C. and Paley,S.M. (1996) HinCyc: a knowledge base of the complete genome and metabolic pathways of *H. influenzae*. In States,D.J., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R. (eds), *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 116–124.
2. Karp,P., Krummenacker,M., Paley,S. and Wagg,J. (1999) Integrated pathway/genome databases and their role in drug discovery. *Trends Biotechnol.*, **17**, 275–281.