The MetaCyc Metabolic Pathway Database

Peter D. Karp

November 12, 2002

Bioinformatics Research Group, SRI International, 333 Ravenswood Ave EK207, Menlo Park, CA 94025; pkarp@ai.sri.com.

To Appear in *Metabolic Engineering*

Portions of this work were published in *Nucleic Acids Research* 30:59–61 2002, reprinted by permission of Oxford University Press.

Abstract

MetaCyc is a metabolic-pathway database (DB) that describes 445 pathways and 1115 enzymes occurring in 158 organisms. As a general reference source on metabolic pathways, MetaCyc serves several purposes in metabolic engineering. It provides an encyclopedic listing of enzymes and their properties to allow a metabolic engineer to find an enzyme whose characteristics solve an engineering problem at hand. It also serves as a reference pathway DB for predicting the metabolic pathway complement of an organism from its genome. Elucidating the pathway network of a new organism is the logical first step in metabolic engineering. MetaCyc is a review-level DB 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. MetaCyc is available through the WWW at `http://MetaCyc.org/`, and is available for local installation as a binary program for the PC and the Sun workstation, and as a set of flat files. Contact metacyc-info@ai.sri.com for information on obtaining a local copy of MetaCyc.

1 Introduction

The MetaCyc database (DB) is an online reference source for metabolic data [1]. It describes metabolic pathways, reactions, enzymes, and substrate and product compounds. MetaCyc is a review-level DB in that a given entry in MetaCyc often integrates information from multiple literature sources.

MetaCyc also is a general reference source on metabolic pathways for the scientific community. It serves as a reference pathway DB for the prediction of the pathway complement of an organism from its annotated genome. Finally, MetaCyc is used as an aid in teaching biochemistry, as well as a resource for metabolic engineering.

MetaCyc aims to provide a smorgasbord 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 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 the EcoCyc DB [2]. 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 the role of MetaCyc in metabolic engineering, the types and scope of the data within MetaCyc, and the Pathway Tools software used to query MetaCyc. The article then compares MetaCyc with the Kyoto Encyclopedia of Genes and Genomes (KEGG).

1.1 Definitions

The following definitions are used in this chapter.

Pathway/Genome Database (PGDB). A database that describes the genome of an organism (its chromosome(s), genes, and genome sequence); the product of each gene; the biochemical reaction(s) catalyzed by each gene product; the substrates of each reaction; and the organization of reactions into pathways.

The ontology (schema) of PGDBs defines 10 major classes of entities: genes, proteins, reactions, pathways, small molecules, chromosomes, RNAs, transcription units, promoters, and DNA binding sites. The ontology defines physical aspects of these entities (such as the subunit structure, pI, and molecular weight of proteins) and functional aspects (such as the reaction catalyzed by an enzyme, and its substrate specificity) [3]. The ontology allows DB curators to store commentary, literature citations, and a history of revisions, in any PGDB object.

The ontology encodes several classification systems, such as a classification of metabolic pathways that organizes MetaCyc pathways into 86 different categories (for example, biosynthetic pathways, biosynthetic pathways for amino acids, and pathways of energy metabolism — see `http://biocyc.org:1555/META/class-subs?object=Pathways`). The ontology also encodes controlled vocabularies, for example, the set of chemical compounds defined as DB objects within MetaCyc defines a controlled list of terms that are used when listing substrates of reactions and modulators of enzyme activity.

Pathway Tools Software. The SRI International (SRI) software used to construct, update, visualize, query, and analyze PGDBs [4]. The three components of the Pathway Tools are

- The **Pathway/Genome Navigator** supports the querying, visualization, and analysis of PGDBs.
- The **Pathway/Genome Editors** supports the interactive updating and refinement of PGDBs.
- The **PathoLogic** pathway-prediction program supports the automated creation of a PGDB and the prediction of the metabolic pathway complement of an organism.

EcoCyc Database: A PGDB for the organism *E. coli* [2]. The majority of the information in EcoCyc is derived from the biomedical literature.

MetaCyc Database: A PGDB containing metabolic data for more than 150 organisms [1].

BioCyc Knowledge Library: The collection of PGDBs at URL `http://BioCyc.org/` is called the BioCyc Knowledge Library. EcoCyc and MetaCyc are component databases within the BioCyc Knowledge Library.

1.2 Significance of MetaCyc for Metabolic Engineering

Metabolic engineering has been defined as "the directed improvement of product formation or cellular properties through the modification of specific biochemical reaction(s) or the modification of new one(s) with the use of recombinant DNA technology" [5]. Metabolic engineering therefore has two key components: understanding the existing biochemical network of an organism, and creating purposeful modifications to that network to alter cellular properties in desired ways. Put another way, the more complete our understanding of the metabolic network of an organism, the more effectively we can modify the network in desired ways. MetaCyc addresses both of these facets of metabolic engineering.

1.2.1 Significance for Understanding of Metabolic Networks

If the same few organisms were always used as hosts for metabolic engineering (e.g., *E. coli* and *B. subtilis*), metabolic engineers could use our existing understanding of the metabolism of those few organisms. However, we expect that a variety of new host organisms will be developed as hosts for metabolic engineering, and that MetaCyc will play a key role in elucidating their metabolic networks. An example is chemical production: new host organisms are needed that have greater solvent tolerance than *E. coli* or *B. subtilis* (so that continuous solvent-based extraction of products can be performed), that can utilize less expensive carbon sources than the glucose required by *E. coli*, and that can better tolerate the heat generated by fermentation (so that expensive cooling is not required). In addition, environmental metabolic-engineering applications such as toxic-waste remediation may be more efficient if they can utilize organisms that are already adapted to growth where the toxic waste occurs (e.g., in soil). Other environmental applications make use of a population of interacting microbial species. Agricultural metabolic engineering will use a variety of plant species as hosts.

Before a quantitative understanding of the metabolic networks of these organisms can be attained using techniques such as flux-balance analysis [6, 7] and metabolic control analysis [8, 9], a qualitative understanding of the networks is required. By qualitative understanding, we mean an enumeration of the enzymes, reactions, and pathways that are present in the organism, and the substrate-level control properties of these enzymes and pathways.

A qualitative characterization of the metabolic network of a new host for metabolic engineering can most rapidly be obtained by genome sequencing followed by bioinformatics analysis of genes and pathways. Genome sequencing yields the nucleotide sequence of the host DNA. Bioinformatics analysis of the genome predicts the locations of genes within the DNA (using programs such as Glimmer [10]); predicts the functions of many of the genes (using techniques such as sequence-similarity searching and motif searching); and predicts the assignments of gene products to metabolic pathways (using programs such as the PathoLogic program developed by our group [11, 4, 12] — see Section 3). The sequencing capacity of the Department of Energy (DOE) Joint Genome Institute (`www.jgi.doe.gov`) enables an entire microbial genome to be sequenced in several days. Bioinformatics analysis can then determine the partial enzyme and metabolic-pathway complement of the organism within several more days. In other words, the qualitative characterization of the biochemical network of an organism that used to require on

the order of 100 person-years of biochemical experimentation can now be done within about 1 or 2 weeks in the best case.

The characterization of a metabolic network using PathoLogic and MetaCyc is both faster and more accurate than manual approaches. Speed is enhanced because the process is automated — a pathway prediction can usually be performed in less than a day, whereas a manual assignment of enzymes to pathways can require weeks. Automated processing is important because the pathway prediction process often must be repeated several times as the genome itself is refined due to accumulation of new sequence data, and the re-computation of gene positions and gene functions. One would not want to repeat a manual pathway prediction several times if each prediction requires weeks of effort. Automated processing also is more accurate because it can consider a much larger set of potential pathways than a person is likely to consider. For example, SRI's application of PathoLogic to the genome of *H. pylori* predicted the presence of 26 pathways that were not detected in a manual analysis of the genome performed by *H. pylori* experts [12].

Note also that the complex interconnections within the metabolic network demand that we characterize the network in its entirety, not just the portion of the network that we think may be relevant to a given engineering problem. Attempts to engineer a given pathway can be confounded by its interactions with other connected pathways that may compensate for engineered changes.

In summary, MetaCyc serves as a crucial link between genomics and metabolic engineering.

1.2.2 Significance for Modification of Metabolic Networks

The modification of a metabolic network through genetic engineering involves (a) inserting a new enzyme or pathway into an organism, (b) replacing an existing enzyme or pathway with a substitute, or (c) removing an enzyme or pathway. Cases (a) and (b) both involve the 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. In most of these cases, the likelihood of finding the right enzyme or pathway is enhanced by the existence of MetaCyc (excepting the case of kinetic information, which MetaCyc does not contain). MetaCyc provides a searchable encyclopedia of enzymes and pathways that lists not only the catalytic function of each enzyme, but also its 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).

Bailey describes metabolic-engineering case studies in which heterologous proteins are introduced into cells to alter their metabolism [13]. He writes that "No universal principles have emerged from metabolic engineering research to guide the choice of the next useful genetic alteration... there is no substitute for knowledge of the pathways involved, their regulation, and their kinetics" [13, p.#1673].

2 The MetaCyc Data

The Ocelot object-oriented database management system (DBMS) [14] was used to construct MetaCyc. Ocelot was chosen because of the well known expressive advantages of object-oriented DBMSs over relational DBMSs, and because of the excellent support for schema evolution provided by Ocelot, which is needed because the schema undergoes constant revision to meet the requirements of this complex domain.

The MetaCyc schema is defined by a set of classes. For example, the class Reactions defines the properties of all biochemical reactions stored within MetaCyc. A set of DB objects define the entries within the MetaCyc DB. For example, objects of the Reactions class define particular biochemical reactions, and objects of the Pathways class define particular biochemical pathways.

MetaCyc version 6.0 contains a total of 888 classes and 11,015 objects. Table 1 shows the number of objects in 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 *E. coli* because MetaCyc contains all metabolic pathways and enzymes from the EcoCyc DB. The MetaCyc data were gathered from a variety of literature and online sources such as books, journals, and the ENZYME DB [15]. Data were manually gathered from the literature, and in some cases are automatically extracted from online sources.

The MetaCyc schema is described in more detail in [16, 17, 3]. In brief, the methanogenesis pathway shown in Figure 1 is represented within MetaCyc as shown in Figure 2. Each box in Figure 2 represents a single object in MetaCyc. Each line in Figure 2 represents a link between two objects in MetaCyc (each of these links are in fact bidirectional). For example, the DB object representing the pathway contains a slot (attribute) whose values are the unique identifers of the three reactions in the pathway, which establish links from the pathway to each reaction. The reactions in turn are linked to objects that represent the enzymes that catalyze each reaction, via an intermediary object (not shown) called an enzymatic-reaction. Each enzyme in this pathway is active as a multimer, and the subunit structure of each enzyme is encoded within MetaCyc by a link from the object describing each multimer to objects describing the monomers from which it is formed. Those monomers in turn are linked to objects that represent the genes that encode them. Reactions are linked to objects that represent the small-molecule reaction substrates (shown only for the bottom reaction). Each object has additional slots that define properties such as the cofactors, activators, and inhibitors to which an enzyme is sensitive, and the chemical structure of a metabolite.

2.1 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.

MetaCyc groups together related *pathway variants* into a common class of pathways. For

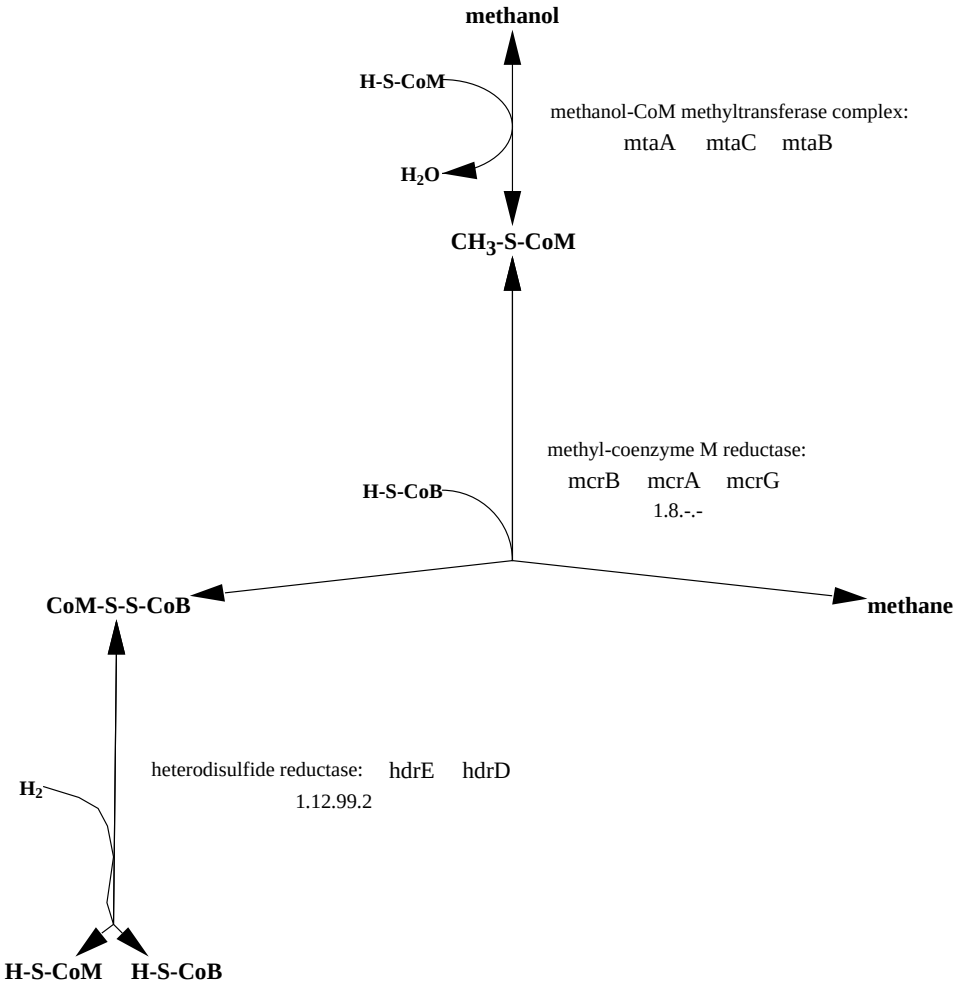

Figure 1: The MetaCyc pathway for methanogenesis from methanol.

r

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 but have similar biochemical functions.

2.2 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 DB. But, because many known reactions have not been captured by the Enzyme Commission (EC) system to date, 689 reactions in MetaCyc do not have an assigned EC number. The DB also contains thousands of objects representing individual metabolites, of which 1860 have chemical structures.

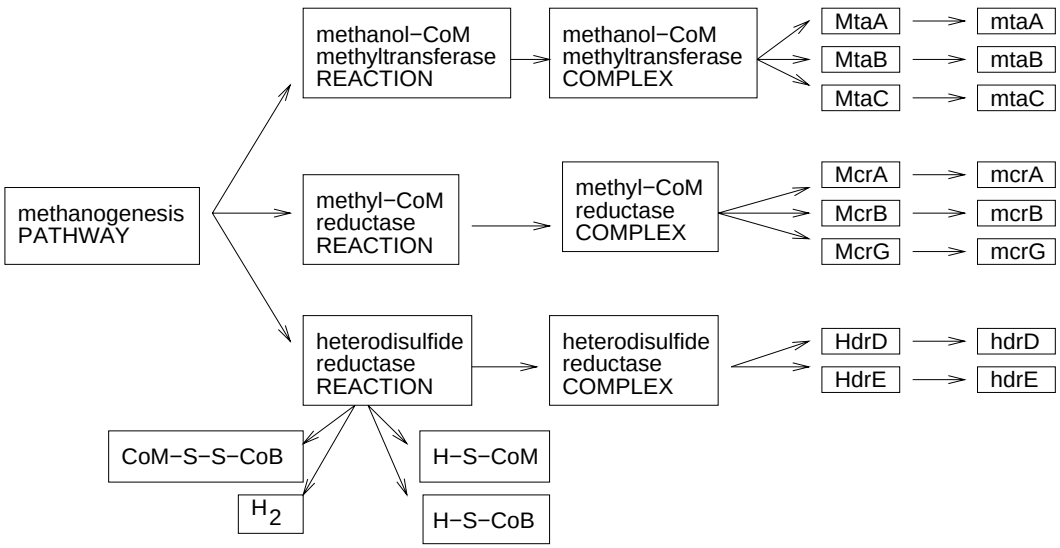

Figure 2: The MetaCyc database representation of the pathway for methanogenesis from methanol.

The number of reactions in MetaCyc that are not assigned to any metabolic pathway is 2232. Some of those reactions are not reactions of small-molecule metabolism. For example, the EC system includes macromolecule reactions involved in transcription, translation, DNA replication, cell signaling, and protein modification. Other reactions may simply not be part of any known pathway. Still other reactions may be part of a known pathway that has not yet been entered into MetaCyc.

The number of MetaCyc reactions that are part of more than one MetaCyc pathway is 392, indicating that there is "cross talk" between pathways in the sense that different pathways reuse the same reaction. For example, glycolysis and gluconeogenesis have multiple reactions in common, and MetaCyc contains eight different variants of the TCA cycle pathway that each share many reactions in common.

The number of MetaCyc reactions that have multiple isozymes defined for them is 252. Some of those isozymes will be from the same species; in other cases the isozymes will be from different species.

2.3 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 DB.

175 of the enzymes in MetaCyc are multifunctional, either because they contain multiple active sites, or because a single active site has loose substrate specificity.

580 MetaCyc enzymes are monomers, and 547 of the enzymes are multimers. Of the 547 multimers, 428 are homomultimers (containing multiple copies of a single subunit) and 119 are heteromultimers.

2.4 Database Links

MetaCyc contains URL-based links to the PIR protein-sequence DB [18], to the ENZYME DB (which links to SWISS-PROT) [15], and to PubMed.

3 Pathway Tools Software

The MetaCyc data reside within the same software environment used for EcoCyc: the Pathway Tools [4]. All the visualization and query tools available for EcoCyc are also available for MetaCyc. For example, the Pathway/Genome Navigator component of Pathway Tools allows users to 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 by 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 DB to predict the metabolic-pathway complement of an organism from its genome [12]. PathoLogic assesses the evidence for the presence of each MetaCyc pathway in the organism under analysis, and creates a new Pathway/Genome DB that models the pathways and genome of that organism. Twelve such bacterial DBs are available at `http://BioCyc.org/`.

4 Comparison of MetaCyc and KEGG

We compare MetaCyc to the KEGG DB [19] 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. For example, the KEGG pathway called "methionine metabolism" combines pathways for the biosynthesis of methionine, charging of methionyl-tRNA, and conversion of methionine

to other compounds such as N-formyl-methionine. The smaller pathways in MetaCyc are advantageous for several reasons. First, the smaller pathways correspond more closely to biologically meaningful units — meaningful in the sense that they correspond to a single biological function, they are regulated as a unit, and they tend to be transferred to other species as a unit. By defining smaller pathways, MetaCyc is able to record and communicate exactly what sequences of reactions are found together in a specific species. By defining connected clusters of individual pathways, called superpathways, MetaCyc does allow the user to view 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. Within the large pathways defined by KEGG, it is impossible for the user to tell which subnetworks correspond to distinct biological units, nor in which species these units have been elucidated experimentally.

MetaCyc pathways are labeled with the name(s) of 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 experimentally 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 does contain cofactor data, although because those data are associated with KEGG reactions rather than with KEGG enzymes, it is difficult to be sure for which proteins from which species the cofactor requirement was experimentally elucidated.

5 Distribution

MetaCyc is available in four forms:

- It is accessible online through the WWW at `http://MetaCyc.org/`. The WWW DB search page is at `http://BioCyc.org:1555/server.html`. The WWW form of access supports a subset of the GUI functionality of the X-windows and PC versions.
- An X-windows version of MetaCyc for the Sun workstation bundles together the Pathway/Genome Navigator software with the MetaCyc DB. This version of MetaCyc can be run as both a client X-windows application, and as a WWW server on a user's intranet.
- A PC version of MetaCyc bundles together the Pathway/Genome Navigator software with the MetaCyc DB. This version of MetaCyc can be run as both a client window application, and as a WWW server on a user's intranet.
- A flatfile version of MetaCyc is available for global analyses.
- An open-source software toolkit is available for loading the MetaCyc flatfiles into a relational DBMS; see URL `biospice.org` (registration required) and search for the BioSPICE Database Warehouse.

All four forms of access are free to academic institutions for research use (see URL `http://BioCyc.org/download.shtml` for information on obtaining MetaCyc). A fee applies to commercial use. The BioCyc WWW site provides background information about the DBs and software, and access to the other publications by SRI's Bioinformatics Research Group.

Acknowledgments

Cindy Kreiger contributed comments on this manuscript. This work was supported by grant 1-R01-GM65466-01 from the National Institutes of Health (NIH) National Institute of General Medical Sciences. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

[1] P.D. Karp, M. Riley, S. Paley, and A. Pellegrini-Toole. The MetaCyc database. *Nuc Acids Res*, 30(1):59–61, 2002.

[2] P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, S. Paley, and A. Pellegrini-Toole. The EcoCyc database. *Nuc Acids Res*, 30(1):56–8, 2002.

[3] P. D. Karp. An ontology for biological function based on molecular interactions. *Bioinformatics*, 16(3):269–285, 2000.

[4] P.D. Karp, S. Paley, and P. Romero. The Pathway Tools Software. *Bioinformatics*, 18:S225–S232, 2002.

[5] G.N. Stephanopoulos, A.A. Aristidou, and J. Nielsen. *Metabolic Engineering: Principles and Methodologies*. Academic Press, San Diego, CA, 1998.

[6] J.S. Edwards, R. Ramakrishna, C.H. Schilling, and B.O. Palsson. Metabolic flux balance analysis. In *Metabolic Engineering*, pages 13–57. Marcel Deker, 1999.

[7] A. Varma and B.O. Palsson. Metabolic flux balancing: Basic concepts, scientific and practical use. *Bio/Technology*, 12:994–8, 1994.

[8] H. Kacser and J.A. Burns. The control of flux. *Symposium Society of Experimental Biology*, 27:65–104, 1973.

[9] R. Heinrich and T.A. Rapoport. A linear steady-state treatment of enzymatic chains. *European Journal Biochemistry*, 42:89–95, 1974.

[10] A.L. Delcher, D. Harmon, S. Kasif, O. White, and S.L. Salzberg. Improved microbial gene identification with GLIMMER. *Nuc Acids Res*, 27:4636–41, 1999.

[11] P.D. Karp, C. Ouzounis, and S.M. Paley. HinCyc: A knowledge base of the complete genome and metabolic pathways of *H. influenzae*. In D.J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, editors, *Proceedings of the Fourth International Conference*

Metabolic Pathways	445
Reactions	4218
Enzymes	1115
Compounds	2335
Citations	2381

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

on Intelligent Systems for Molecular Biology, pages 116–24, Menlo Park, CA, 1996. AAAI Press.

[12] S. Paley and P.D. Karp. Evaluation of computational metabolic-pathway predictions for *H. pylori*. Bioinformatics, 18(5):705–714, 2002.

[13] J.E. Bailey. Toward a science of metabolic engineering. *Science*, 252:1668–1675, June 1991.

[14] P.D. Karp, Vinay K. Chaudhri, and Suzanne M. Paley. A collaborative environment for authoring large knowledge bases. *Journal of Intelligent Information Systems*, 13:155–94, 1999.

[15] A. Bairoch. The ENZYME databank in 2000. *Nuc Acids Res*, 28(1):304–305, 2000.

[16] P. Karp and M. Riley. Representations of metabolic knowledge. In L. Hunter, D. Searls, and J. Shavlik, editors, *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, pages 207–215, Menlo Park, CA, 1993. AAAI Press.

[17] P. Karp and S. Paley. Representations of metabolic knowledge: Pathways. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 203–211, Menlo Park, CA, 1994. AAAI Press.

[18] W.C. Barker, J.S. Garavelli, H. Huang, P.B. McGarvey, B.C. Orcutt, G.Y. Srinivasarao, C. Xiao, L.-S.L. Yeh, R.S. Ledley, J.F. Janda, F. Pfeiffer, H.-W. Mewes, J.A. Tsugita, and C. Wu. The protein information resource PIR. *Nuc Acids Res*, 28(1):41–44, 2000.

[19] M. Kanehisa, S. Goto, Shuichi Kawashima, and Akihiro Nakaya. The KEGG databases at GenomeNet. *Nuc Acids Res*, 30:42–46, 2002.

Escherichia coli	173
Salmonella typhimurium	35
Homo. sapiens	31
Sulfolobus solfataricus	20
Pseudomonas	20
Bacillus subtilis	18
Glycine max	18
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
Pseudomonadacea	3
Neisseriaceae	3
Klebsiella aerogenes	3
Rattus norvegicus	3
Methanosarcina barkeri	3
Sinorhizobium meliloti	3

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