50–53 Nucleic Acids Research, 1998, Vol. 26, No. 1

EcoCyc: Encyclopedia of *Escherichia coli* genes and metabolism

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

Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA and [1]Marine Biological Laboratory, Woods Hole, MA 02543, USA

Received September 29, 1997; Accepted October 3, 1997

ABSTRACT

The encyclopedia of *Escherichia coli* genes and metabolism (EcoCyc) is a database that combines information about the genome and the intermediary metabolism of *E.coli*. The database describes 3030 genes of *E.coli*, 695 enzymes encoded by a subset of these genes, 595 metabolic reactions that occur in *E.coli*, and the organization of these reactions into 123 metabolic pathways. The EcoCyc graphical user interface allows scientists to query and explore the EcoCyc database using visualization tools such as genomic-map browsers and automatic layouts of metabolic pathways. EcoCyc can be thought of as an electronic review article because of its copious references to the primary literature, and as a (qualitative) computational model of *E.coli* metabolism. EcoCyc is available at URL http://ecocyc.PangeaSystems.com/ecocyc/

INTRODUCTION

The encyclopedia of *Escherichia coli* genes and metabolism (EcoCyc) is a database (DB) that combines information about the genome and the intermediary metabolism of *E.coli* K-12. The DB describes most known genes of *E.coli*, the enzymes of small-molecule metabolism that are encoded by these genes, the reactions catalyzed by each enzyme, and the organization of these reactions into metabolic pathways. EcoCyc can be viewed as an electronic review article because it is a carefully sifted collection of information drawn largely from (and containing 1650 citations to) the primary literature. The EcoCyc graphical user interface (GUI) allows scientists to query, explore, and visualize the EcoCyc DB. EcoCyc integrates genomic and functional data to allow scientists to investigate a broad range of questions (4).

Among the problems that might be addressed using EcoCyc are the following (some of these tasks are not directly supported by the EcoCyc user interface and would require additional programming).

- EcoCyc is a resource for analysis of microbial genomes. For example, EcoCyc has been used to predict the metabolic pathways of *H.influenza* (9) and of *H.pylori* (11).
- Because of its links to sequence DBs such as Swiss-Prot, EcoCyc can be used to perform function-based retrieval of DNA or protein sequences, for example to prepare datasets for studies of protein structure–function relationships.
- Scientists who study the evolution of metabolism can use EcoCyc to search out examples of duplication and divergence of enzymes and pathways.
- EcoCyc provides a foundation for performing simulations of the metabolism, although it currently lacks the kinetics data used by most simulation techniques.
- The DB has been used as an aid in teaching biochemistry.

This article describes recent enhancements to EcoCyc and how to access EcoCyc. We request that users of EcoCyc cite this article in publications related to its use.

RECENT ENHANCEMENTS

In the past year we supplemented the EcoCyc data with the following 13 new pathways: glutamine utilization; L-serine degradation; glutamate utilization; L-cysteine catabolism; tryptophan utilization; 2-phenylethylamine degradation; enterobactin synthesis; aerobic electron transfer; anaerobic electron transfer; carnitine metabolism, CoA-linked; carnitine metabolism; pyridine nucleotide cycling; nucleotide metabolism.

We have reorganized the Overview diagram of the *E.coli* metabolic map. The Overview is now available through the WWW at http://ecocyc.PangeaSystems.com/ecocyc/ov.html , and is shown in Figure 1. The new organization reflects the new pathways added to EcoCyc in the past year, and also reflects a new organizing principle: anabolic pathways are drawn on the left side of the diagram, catabolic pathways are drawn on the right side, and energy-producing pathways are drawn in the middle. (Because some metabolic reactions perform more than one role under different metabolic circumstances, we made a choice as to the primary role.) EcoCyc provides several queries that operate on the Overview. Users can highlight objects in the Overview, such as finding compounds by name or substring, finding reactions by EC number, or finding enzymes by name or substring. Users can also highlight enzymes according to their sensitivity to metabolites, such as all enzymes that are inhibited by ATP or that are activated by l-lactate. EcoCyc maintains a

*To whom correspondence should be addressed at +present address: Pangea Systems Inc., 4040 Campbell Avenue, Menlo Park, CA 94025, USA. Tel: +1 510 628 0100; Fax: +1 510 628 0108; Email: pkarp@pangeasystems.com

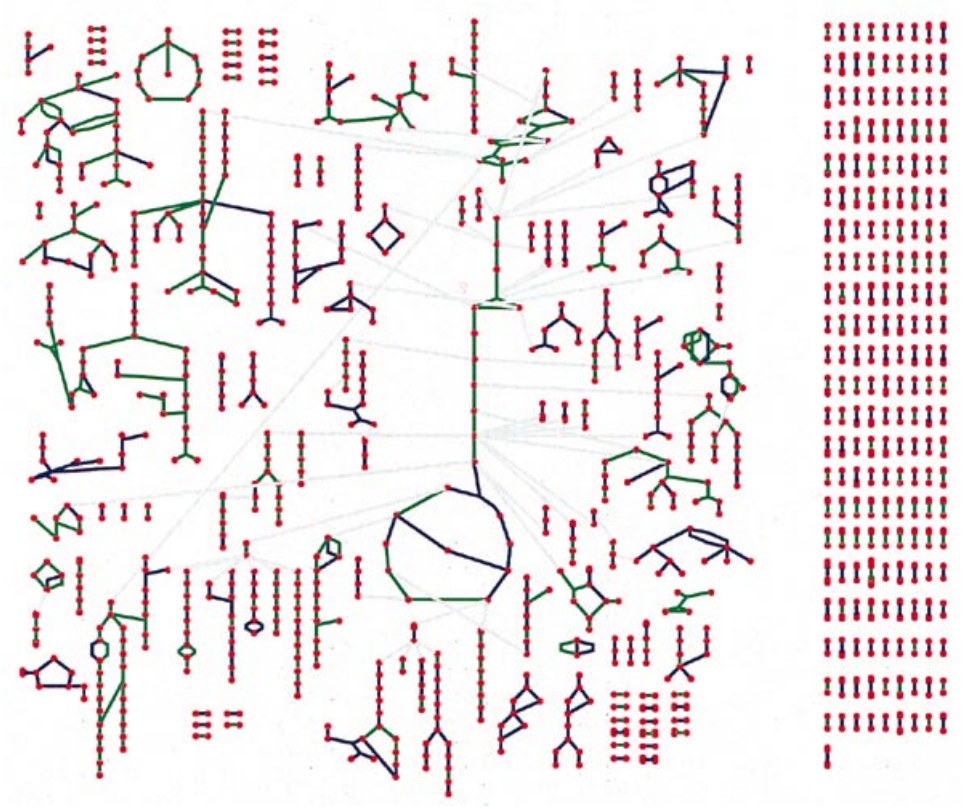

Figure 1. Guide to the EcoCyc schema. This version of the Overview diagram shows a comparison between the full metabolic network of *E.coli* and the predicted metabolic network of *H.influenza*. Each circle represents a single metabolite. Each blue line represents a metabolic reaction that occurs in *E.coli* only; each green line represents a reaction that occurs in both *E.coli* and *H.influenza*. Each grey line connects two dots that represent the same metabolite.

history of the highlighting operations, so users can undo or redo their highlighting queries. Highlighting operations are not supported in the WWW version of EcoCyc.

A new visualization within EcoCyc pathway displays shows the distribution of genes that encode the enzymes of a pathway. The visualization consists of a small circle representing the chromosome with tick marks drawn for each gene in the pathway. When the user moves the mouse pointer over a tick mark, EcoCyc flashes the name of the gene, and highlights the arrow within the pathway drawing for the reaction(s) that are catalyzed by the gene product.

We have also enhanced pathway visualizations to depict polymerization steps (see Fig. 2). We use a dashed line to indicate that two compound names are, in certain situations, meant to represent the same chemical species. For example, most textbooks depict saturated fatty acid elongation as a spiral, where each turn of the spiral adds two carbons to the backbone. Our representation shows the pathway as a cycle, using generic rather than specific names for the compounds involved. At the 'beginning' of the cycle is acyl$_N$-ACP, which undergoes several reactions producing acyl$_{N+2}$-ACP. A dashed line is drawn between these two names to indicate that the N+2 species becomes the N species for the next iteration of the cycle. We also use the dashed line when showing equivalence between a specific name for a compound (such as a starting or ending compound for a series of polymerization reactions) and the generic form. Using this scheme, we can compactly represent polymerization pathways as cycles of generic compounds, with specific compounds as inputs and/or outputs.

EcoCyc now contains descriptions of 79 tRNAs. Each tRNA is represented as a distinct object within the DB, and is linked to the EcoCyc object that represents the gene for the tRNA. 33 tRNA synthetases, and the associated charging reactions, are also encoded as EcoCyc objects, where the tRNA objects are substrates in these reactions. Additional substrates include the charged tRNAs, which are also represented as distinct objects within the DB.

The reactions of two-component signal transduction systems in *E.coli* have been added to EcoCyc. About 22 signal transduction systems are in *E.coli* involving at least two gene products each. These types of regulatory reactions have counterparts in eukaryotic organisms and are in this sense housekeeping functions with ancient common ancestors.

The two components, the sensor protein and the response regulator protein, interact to convert an environmental signal (either internal or external) into regulation of relevant gene expression. Although the systems differ, generally the sensor protein becomes phosphorylated when stimulated by a specific condition such as lack of oxygen or shortage of nitrogen. The phosphorylated sensor transfers the phosphate to a regulator protein, which then transfers the phosphate to another of its amino acid residues internally, accompanied by an allosteric change of the regulator protein. The altered regulator is then active as a transcriptional activator. Although signal transduction systems of

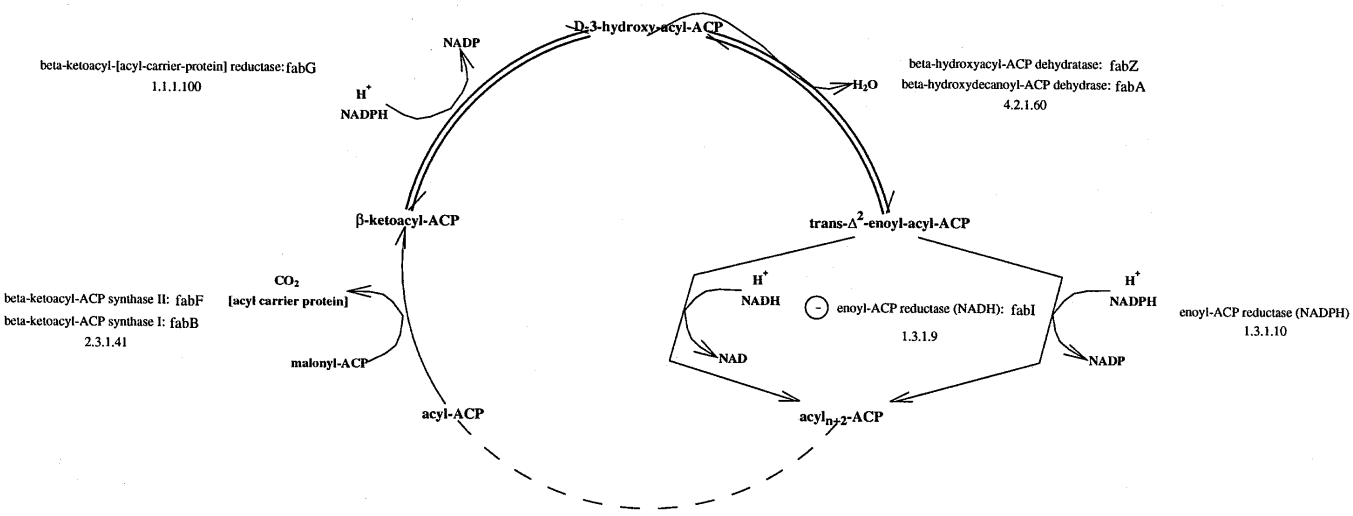

Figure 2. An EcoCyc drawing of the pathway for elongation of saturated fatty acids.

E.coli exhibit broad similarities, there are at least three classes with different modes of action corresponding to different amino acid sequence domains. The reactions of these systems often take the form of a cascade of sequential events. The reactions are, in some sense, in the realm of macromolecule metabolism, because proteins act both as substrates and products in reactions that modify the covalent composition of the proteins. We represent the functions of two-component signal transductions, such as phosphorylation events, as reactions.

EcoCyc reactions are now linked to three other metabolic DBs: ENZYME (1), WIT (10), and Ligand. Table 1 summarizes all the DBs to which EcoCyc is linked.

By the time this article appears, we expect that data from the full genomic sequence of *E.coli* (2) will be incorporated into EcoCyc. We plan to create EcoCyc objects for all *E.coli* genes, and to incorporate the map positions determined the Blattner group.

Table 1. The biological DBs to which objects in different EcoCyc classes are linked, e.g., EcoCyc genes are linked to the CGSC DB and to GenBank

Class	Linked databases
Genes	Coli Genetic Stock Center, GenBank
Polypeptides	Expasy Swiss-Prot, NCBI Swiss-Prot, PDB, Swiss-Model
Citations	PubMed
Reactions	ENZYME, WIT, Ligand

THE EcoCyc GRAPHICAL USER INTERFACE

The EcoCyc GUI (3) provides graphical tools for visualizing and navigating through an integrated collection of metabolic and genomic information (its retrieval capabilities are described in ref. 8). For each type of biological object in the EcoCyc DB, the GUI provides a corresponding visualization tool. These tools dynamically query the underlying DB. Most display algorithms are parameterized to allow the user to select the visual presentation of an object that is most informative. For example, the algorithms that produce automatic layouts of metabolic pathways can suppress the display of enzyme names or side-compound names; they can also draw chemical structures for the compounds within a pathway. More details on the display algorithms can be found in ref. 5.

THE EcoCyc DATA

The EcoCyc data are stored within a frame knowledge representation system (FRS) called Ocelot. FRSs use an object-oriented data model. FRSs organize information within classes: collections of objects that share similar properties and attributes. Table 2 shows the current size of several EcoCyc classes. These statistics pertain to EcoCyc version 3.8.

The current scope of metabolic information within EcoCyc is intermediary metabolism only; EcoCyc does not cover macromolecule metabolism such as DNA replication or repair, nor transcription, nor translation. It does describe tRNA charging.

For more information on the contents of EcoCyc and the data validation procedures we employ, see ref. 8; the EcoCyc schema is defined in ref. 7. The retrieval operations supported by the DB are described in refs 3 and 8. The EcoCyc software architecture is described in ref. 6.

Table 2. The number of objects in several EcoCyc classes

Reactions	595
Enzymes	695
Pathways	123
Genes	3030
tRNAs	79
Compounds	1296

DISTRIBUTION

EcoCyc is available under license from Pangea Systems via the Internet in three forms: (i) a program for the Sun workstation

bundles together the EcoCyc GUI and the EcoCyc DB; (ii) EcoCyc is accessible online through the WWW (this version supports a subset of the GUI functionality of the Sun-workstation version); (iii) the EcoCyc DB alone is available in a limited fashion as a set of flat files. Access is free to academic institutions for research use; a fee applies to other forms of use.

The EcoCyc WWW pages describe all three types of access to EcoCyc; they also provide links to the EcoCyc User's Guide, to detailed documentation of the EcoCyc schema, and to all publications produced by the EcoCyc project. The URL for the EcoCyc home page is http://ecocyc.PangeaSystems.com/ecocyc/

ACKNOWLEDGEMENTS

This work was supported by grant 1-R01-RR07861-01 from the National Center for Research Resources, and by grant R29-LM-05413-01A1 from the National Library of Medicine. 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 Bairoch,A. (1996) *Nucleic Acids Res.*, **24**, 221–222.
2 Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T. Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) *Science*, **277**, 1453–1462.
3 Karp,P. (1996) The EcoCyc User's Guide. ftp://ftp.ai.sri.com/pub/papers/karp-ecocyc-guide.ps.Z
4 Karp,P. and Mavrovouniotis,M. (1994) *IEEE Expert*, **9**, 11–21.
5 Karp,P. and Paley,S. (1995) In Lim,H., Cantor,C. and Robbins,R. (eds), *Proceedings of the Third International Conference of Bioinformatics and Genome Research*. World Scientific Publishing Co., pp. 225–238. See also ftp://ftp.ai.sri.com/pub/papers/karp-bigr94.ps.Z
6 Karp,P. and Paley,S. (1996) *J. Comp. Biol.*, **3**, 191–212.
7 Karp,P. and Riley,M. (1996) Guide to the EcoCyc schema. ftp://ftp.ai.sri.com/pub/papers/karp-ecocyc-schema.ps
8 Karp,P., Riley,M., Paley,S., Pellegrini-Toole,A. and Krummenacker,M. (1997) *Nucleic Acids Res.*, **25**, 43–50.
9 Karp,P.D., Ouzounis,C. and Paley,S.M. (1996) 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.
10 Selkov,E., Galimova,M., Goryanin,I., Gretchkin,Y., Ivanova,N., Komarov,Y., Maltsev,N., Mikhailova,N., Nenashev,V., Overbeek,R. *et al.* (1997) *Nucleic Acids Res.*, **25**, 37–38. See also this issue *Nucleic Acids Res.* (1998), **26**, 43–45.
11 Tomb,J.-F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al.* (1997) *Nature*, **388**, 539–547.