Nucleic Acids Research, 1999, Vol. 27, No. 1 55–58

Eco Cyc: Encyclopedia of *Escherichia coli* genes and metabolism

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

Pangea Systems Inc., 4040 Campbell Avenue, Menlo Park, CA 94025, USA and [1]Marine Biological Laboratory, Woods Hole, MA 02543, USA

Received October 2, 1998; Revised October 8, 1998; Accepted October 14, 1998

ABSTRACT

The EcoCyc database describes the genome and gene products of *Escherichia coli*, its metabolic and signal-transduction pathways, and its tRNAs. The database describes 4391 genes of *E.coli*, 695 enzymes encoded by a subset of these genes, 904 metabolic reactions that occur in *E.coli*, and the organization of these reactions into 129 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 has many references to the primary literature, and is 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 describes all known genes of *E.coli* K-12, 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 also describes *E.coli* signal-transduction pathways, and *E.coli* tRNAs. EcoCyc can be viewed as an electronic review article because it is a carefully sifted collection of information drawn largely from (and containing 1834 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 (1).

EcoCyc is employed for the following tasks by the scientific community. (i) EcoCyc is a resource for analysis of microbial genomes at the level of individual genes and entire pathways. Because the *E.coli* genome has a high fraction of genes whose functions were determined experimentally, it is an accurate reference for inferring gene function by sequence similarity. The metabolic pathways within EcoCyc have been used to predict the metabolic pathways of *Haemophilus influenzae* (2) and of *Helicobacter pylori* (3). (ii) 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. (iii) Scientists who study the evolution of metabolism can use EcoCyc to search out examples of duplication and divergence of enzymes and pathways. (iv) EcoCyc provides a foundation for performing simulations of the metabolism, although it currently lacks the kinetics data used by most simulation techniques. (v) The DB is 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.

Two new versions of EcoCyc were released in 1998: version 4.0 (released in April, 1998) and version 4.5 (released in September, 1998).

Table 1. The number of objects in EcoCyc version 4.5

Metabolic Pathways	129
Signaling Pathways	20
Reactions	904
Enzymes	695
Genes	4391
tRNAs	79
Compounds	1854
Citations	1834

THE EcoCyc DATA

The EcoCyc data are stored within a frame knowledge representation system (FRS) called Ocelot (4,5). FRSs use an object-oriented data model that organizes information within classes: collections of objects that share similar properties and attributes. Table 1 shows the current size of several EcoCyc classes.

For more information on the contents of EcoCyc and the data validation procedures we employ, see ref. 6. The retrieval operations supported by the DB are described in ref. 6 and in the EcoCyc User's Guide at http://ecocyc.PangeaSystems.com/ecocyc/doc/ecocyc-uguide/paper.html . The EcoCyc software architecture is described in ref. 4.

*To whom correspondence should be addressed. Tel: +1 650 614 7066; Fax: +1 650 324 9313; Email: pkarp@pangeasystems.com

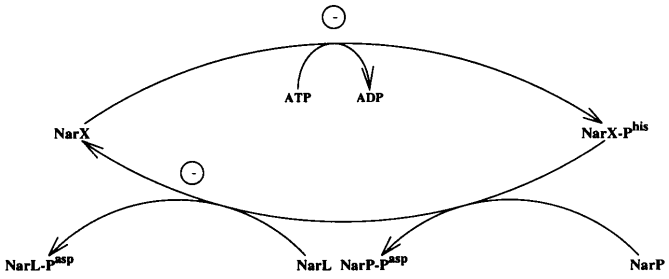

Figure 1. The NarX nitrate/nitrite-dependent two-component regulatory system. The diagram shows a number of linked phosphorylation reactions of the NarX, NarL and NarP proteins. The circled minus signs indicated that inhibitors of those reactions are known; clicking on the minus sign will list the inhibitors.

Two-component signal-transduction pathways

Pathways of two-component signal transduction in *E.coli* were added to EcoCyc version 4.0. The two components, the sensor protein and the response regulator protein, interact to convert an environmental signal (either internal or external) into regulation of gene expression. In EcoCyc version 4.5 we extended the pathway-layout capabilities of EcoCyc to support drawing of signal-transduction pathways, as shown in Figure 1.

Incorporation of the full *E.coli* nucleotide sequence

Version 4.0 of EcoCyc was the first to contain the full *E.coli* genome as determined by the Blattner laboratory (7). The main challenge we faced in incorporating the Blattner-lab data into EcoCyc was how to add new information from the full genomic sequence without losing the unique information about 3030 *E.coli* genes that was already contained within EcoCyc version 3.7. The EcoCyc gene objects contained information such as gene-name synonyms, unique IDs in use by external DBs that linked to EcoCyc (such as Swiss-Prot), and links to polypeptide objects within EcoCyc that describe the products of these genes.

We therefore proceeded to determine as many correspondences as possible between genes within the Blattner-lab GenBank entry (accession number U00096) and previously existing EcoCyc genes. We did so by matching the names of genes within U00096 and the names of EcoCyc genes; we confirmed those matches by cross-checking the Swiss-Prot IDs included in U00096 with Swiss-Prot links stored within EcoCyc. This procedure found 2257 matches. For each gene we imported from U00096 the starting and ending base-pair position of the gene within the chromosome, the unique ID (b-number) assigned to the gene by the Blattner lab, and synonyms for the gene name. We compared the product name assigned in U00096 with that assigned within EcoCyc.

1874 genes in U00096 did not match any EcoCyc gene; we therefore created 1874 new gene objects in EcoCyc containing the same imported information, plus the gene product assigned by the Blattner lab. Centisome positions for all genes were recomputed from the base-pair positions from U00096. 552 genes within EcoCyc did not match any gene from U00096. These genes were retained within EcoCyc; they represent genes reported in the literature that had not been associated with a particular ORF within the *E.coli* genome. Over time we have identified many correspondences between those genes and *E.coli* ORFs, and we have merged those corresponding genes together, so that now only 279 of those 552 genes remain without a chromosomal location.

Whenever a gene-merging event is performed within EcoCyc, the event is recorded in the resulting gene object. Some current EcoCyc gene objects were derived from multiple merging operations as we determined that genes reported under different names in the literature were in fact one and the same gene. For example, the two history entries shown below for the gene *glnU* reflect two merging events undergone by this gene. The earlier history entry indicates the merging of a gene in U00096 with an EcoCyc gene whose internal id is EG30028; the resulting gene was later merged with the EcoCyc gene whose id was G791 when we determined that these two objects within EcoCyc described the same gene.

7/10/1998 Merged genes G791/trnA and EG30028/glnU.
10/20/1997 Gene b0670 from Blattner lab Genbank (vM52) entry merged into EcoCyc gene EG30028.

The nucleotide sequence of *E.coli* genes is accessible to EcoCyc users in two ways. Within a gene window the user may click on the button **Show Sequence** to retrieve the nucleotide sequence of that gene from U00096 (EcoCyc does not currently provide access to non-coding DNA). Or the user may click on the button **Query Genbank** to query the NCBI Entrez server for all Genbank entries for *E.coli* containing a gene with the same name as the current gene.

The evolving annotation of the *E.coli* genome

We seek to make the annotation of the *E.coli* genome within EcoCyc as complete and up-to-date as possible based on new functional characterizations of *E.coli* genes in the literature, and based on an ongoing sequence analysis of the *E.coli* genome by the EcoCyc project.

EcoCyc is very careful to distinguish genes whose functions have been determined experimentally, from those whose functions have been determined through sequence analysis. The names of gene products whose functions have been determined through sequence analysis always contain the word ‘putative’ and very occasionally, for a high certainty hit, ‘probable’. The level of assurance for each functional assignment is given in the following way. If an ORF sequence shows similarity to a number of hydrolases, all of which act on sugars, the ORF is identified as, for instance, a ‘putative sugar hydrolase’. In other cases, an assignment may be ‘putative amidotransferase’, ‘putative aminotransferase’, or ‘putative formyl acetyltransferase’. More often a degree of uncertainty is signified by confining the assignment to a general class such as ‘putative transferase’, or sometimes only as ‘an enzyme’ if the ORF can be identified as an enzyme, but not what kind of enzyme. The same gradations are used for other types of gene products such as regulators, transport components and RNAs.

Some assignments of putative function have been made on the basis of similarity among paralogous sequences of proteins within *E.coli* (8). The advantage of using paralogous groups is that within one genome even when sequence similarity within these sets is weak, functions are the same or closely related.

In the initial annotation of the *E.coli* genome published by Blattner *et al*. in September 1997, 38% of the open reading frames had no attributed function. In EcoCyc 4.5, there are 1400 ORFs

Table 2. A taxonomy of genes according to the physiological role of the gene product

SMALL MOLECULE METABOLISM	886
Amino acid biosynthesis	114
leucine	7
isoleucine/valine	17
alanine	2
histidine	9
chorismate	8
tryptophan	9
tyrosine	3
phenylalanine	3
cysteine	5
serine	3
glycine	1
methionine	8
threonine	5
lysine	10
asparagine	3
aspartate	1
proline	3
arginine	10
glutamine	5
glutamate	2
Biosynthesis of cofactors, carriers	136
fatty acid and phosphatidic acid biosynth	24
enterochelin	6
cobalamin	5
heme, porphyrin	16
menaquinone, ubiquinone	16
thioredoxin, glutaredoxin, glutathione	9
riboflavin	5
thiamin	10
pyridine nucleotide	7
pyridoxine	4
pantothenate	4
molybdopterin	9
lipoate	2
folic acid	9
biotin	9
acyl carrier protein (ACP)	0
biotin carboxyl carrier protein (BCCP)	1
Fatty acid biosynthesis	0
Nucleotide biosynthesis	31
pyrimidine ribonucleotide biosynthesis	10
purine ribonucleotide biosynthesis	21
Global functions	64
global regulatory functions	55
ATP-proton motive force interconversion	9

Each line in the figure indicates a single class. A new level of indentation indicates a subclass of the class above. The numbers in the right hand column indicate the number of genes within each class.

Table 3. A taxonomy of genes continued

SMALL MOLECULE METABOLISM	886
Central intermediary metabolism	183
salvage of nucleosides and nucleotides	19
2-prime-deoxyribonucleotide metabolism	9
sulfur metabolism	13
misc. glucose metabolism	3
nitrogen metabolism	2
oligosaccharides	0
amino sugars	14
nucleotide interconversions	11
sugar-nucleotide biosynthesis, conversions	9
misc. glycerol metabolism	0
nucleotide hydrolysis	2
gluconeogenesis	5
amino acids	1
polyamine biosynthesis	9
Entner-Douderoff	3
glyoxylate bypass	3
pool of unassigned individual reversible reactions	52
non-oxidative branch, pentose pwy	8
phosphorus metabolism	20
Energy transfer	29
electron transport	29
Energy metabolism, carbon	166
fermentation	23
anaerobic respiration	67
aerobic respiration	33
oxidative branch, pentose pwy	3
TCA cycle	18
pyruvate dehydrogenase	5
glycolysis	17
Degradation	163
fatty acids	12
amines	3
amino acids	22
carbon compounds	126

Table 4. A taxonomy of genes continued

MACROMOLECULE METABOLISM	376
Macromolecule degradation	74
degradation of DNA	26
degradation of RNA	13
degradation of polysaccharides	3
degradation of proteins, peptides, glycopept.	32
Macromolecule synthesis, modification	170
phospholipids	11
lipopolysaccharides	0
glycoproteins	0
lipoproteins	1
polysaccharides - (cytoplasmic)	7
RNA synthesis, modification, DNA transcrip.	28
proteins & peptides - translation and modification	34
DNA - replication, repair, restr./modifn.	89
Basic proteins	6
basic proteins - synthesis, modification	6
aa-tRNAs	126
amino acyl tRNA syn; tRNA modification	47
tRNA	79
STRUCTURAL ELEMENTS	271
Cell Exterior	103
surface polysaccharides/antigens	43
surface structures	60
Cell envelope	72
membrane, inner	0
membrane, outer	31
murein sacculus, peptidoglycan	41
Ribosome constituents	96
ribosomes - maturation and modification	14
ribosomal proteins - synthesis, modification	57
ribosomal and stable RNAs	25

with no attributed function (32%), and 939 genes whose attributed function is marked as putative (21%).

Because EcoCyc combines the *E.coli* genome with experimentally derived information about *E.coli* gene products, we can assess the degree of correspondence between these two bodies of knowledge. For example, we can write a query to EcoCyc to retrieve all enzymes within EcoCyc for which the gene has not yet been determined (in the case of enzymes known to have multiple subunits, we require that none of the subunits have a gene assigned). That list of enzymes is given at URL http://ecocyc.PangeaSystems.com/ecocyc/enzymes.html , and represents a challenge to both experimentalists and bioinformaticians.

EcoCyc taxonomies

The EcoCyc project has developed several taxonomies for the different types of biological information within EcoCyc. It includes the taxonomy of gene products developed by Riley (9), which has been adopted by a number of other genome projects. The most recent version of that taxonomy is shown in Tables 2, 3, 4 and 5. EcoCyc also includes a taxonomy of biochemical pathways, shown in Table 6.

These taxonomies are accessible within EcoCyc for taxonomic querying of EcoCyc objects. For example, the user can easily navigate to the pathways within any of the classes in Table 6, using the query page at URL http://ecocyc.PangeaSystems.com/ecocyc/server.html

Table 5. A taxonomy of genes continued

CELL PROCESSES	457
Cell division	37
Chaperoning	7
Motility chemotaxis	17
Transport of small molecules	251
anions	17
amino acids, amines	58
carbohydrates, organic acids, alcohols	89
cations	61
incorporation metal ions	0
nucleosides, purines, pyrimidines	7
other transport	19
Transport of large molecules	40
uptake of DNA	0
protein, peptide secretion	40
Adaptation	35
adaptations, atypical conditions	20
osmotic adaptation	15
Protection responses	70
radiation sensitivity	7
drug/analog sensitivity	48
cell killing	3
detoxification	12
ELEMENTS OF EXTERNAL ORIGIN	47
transposon-related functions	5
plasmid-related functions Non-bacterial functions	3
colicin-related functions Non-bacterial functions	10
phage-related functions and prophages	29
MISCELLANEOUS	65
not classified	65
ORFs	1400

Table 6. A taxonomy of pathways

Biosynthesis	60
Amino acid biosynthesis	30
Amino acid families	10
Individual amino acids	20
Carbohydrates	1
Cell-structures	7
Surface structures	5
Murein sacculus	2
Cofactors, prosthetic groups, electron carriers	15
Fatty acids and lipids	4
Nucleotides	2
2'-deoxyribonucleotides	0
purines and pyrimidines	2
Ribonucleotides	0
Polyamines	1
Degradation	44
Amino acids, amines	15
Carbon compounds	24
Fatty acids	2
Phosphorus compounds	0
Other	3
Energy metabolism	20
Intermediary metabolism	29
Central intermediary metabolism	21
Network of nucleotide interconversions	4
Nitrogen metabolism	0
Sulfur metabolism	1
Signal-transduction pathways	20

Each line in the figure indicates a single class. A new level of indentation indicates a subclass of the class above. For example, degradation of 'Amino acids, amines' is a subclass of the general category of degradation pathways. The numbers in the right hand column indicate the number of individual EcoCyc pathways within each class.

THE EcoCyc GRAPHICAL USER INTERFACE

The EcoCyc GUI provides graphical tools for visualizing and navigating through an integrated metabolic/genomic DB. For each type of biological object in the EcoCyc DB, the GUI provides a corresponding visualization tool that dynamically queries the underlying DB.

Version 4.5 includes a number of extensions to the query operations of the Overview diagram that displays the full metabolic map of *E.coli*. For example, the Overview can be used to highlight those reaction steps that are activated or inhibited by a specified compound, those steps for which *E.coli* has multiple isozymes, or those steps that are shared or not shared with other organisms for which a metabolic DB is available [such as the HinCyc DB (2)]. The diagram can also be used to visualize gene-expression data.

The EcoCyc pathway-layout algorithms have been extended to draw lines that indicate feedback inhibition and activation of an enzyme by substrates of the pathway containing the enzyme.

DISTRIBUTION

EcoCyc is available under license from Pangea Systems in two forms: (i) EcoCyc is accessible online through the WWW (this version supports a subset of the GUI functionality of the X-windows version). (ii) An X-windows version of EcoCyc for the Sun workstation bundles together the EcoCyc GUI and the EcoCyc DB.

Access is free to academic institutions for research use; a fee applies to other forms of use. The EcoCyc WWW pages describe both types of access to EcoCyc; they also provide links to the EcoCyc User's Guide, and to the 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. 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. and Mavrovouniotis,M. (1994) *IEEE Expert*, **9**, 11–21.
2 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.
3 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.
4 Karp,P. and Paley,S. (1996) *J. Computat. Biol.*, **3**, 191–212.
5 Karp,P.D., Chaudri,V.K. and Paley,S.M. (1999) *J. Intelligent, Information Syst.*, in press.
6 Karp,P., Riley,M., Paley,S., Pellegrini-Toole,A. and Krummenacker,M. (1997) *Nucleic Acids Res.*, **25**, 43–50.
7 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.
8 Riley,M. and Labedan,B. (1997) *J. Mol. Biol.*, **268**, 857–868.
9 Riley,M. (1993) *Microbiol. Rev.*, **57**, 862–952.