EcoCyc: an encyclopedia of *Escherichia coli* genes and metabolism

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

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

Received August 31, 1995; Revised and Accepted October 17, 1995

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*. It describes 2034 genes, 306 enzymes encoded by these genes, 580 metabolic reactions that occur in *E.coli* and the organization of these reactions into 100 metabolic pathways. The EcoCyc graphical user interface allows query and exploration of the EcoCyc database using visualization tools such as genomic map browsers and automatic layouts of metabolic pathways. EcoCyc spans the space from sequence to function to allow investigation of an unusually broad range of questions. EcoCyc can be thought of as both an electronic review article, because of its copious references to the primary literature, and as an *in silico* model of *E.coli* that can be probed and analyzed through computational means.

INTRODUCTION

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*. It describes the known genes of *E.coli*, the enzymes encoded by these genes, the reactions catalyzed by each enzyme and the organization of these reactions into metabolic pathways. The EcoCyc graphical user interface (GUI) allows query, exploration and visualization of the EcoCyc database. EcoCyc spans the space from sequence to function to allow investigatation of an unusually broad range of questions (1).

This article describes the information contained within EcoCyc and discusses potential uses of EcoCyc. Our aim is to provide a thorough description of the database by discussing the scope of its current contents, the conceptualization employed to structure the database, the sources from which we obtained the EcoCyc data and the procedures used to construct the database and to verify its correctness. The article also describes our software for retrieving and visualizing EcoCyc data.

MOTIVATIONS

EcoCyc can be viewed as an electronic review article, because it is a carefully sifted collection of information drawn largely from (and containing citations to) the primary literature. In this respect its aim differs from that of databases such as GenBank, because GenBank is designed as a repository of primary observations (an electronic mirror of the primary literature). Another difference between EcoCyc and databases such as GenBank is that EcoCyc describes several classes of biological objects (such as proteins, genes and pathways), whereas GenBank describes only nucleic acid sequences. EcoCyc is therefore an electronic reference source on *E.coli*. However, EcoCyc is also designed to facilitate complex computations on genomic and metabolic data and to provide an *in silico* model of *E.coli* that can be probed and analyzed through computational means.

Problems that might be addressed using EcoCyc include the following.

Genomic investigations

Coupled with sequence databases EcoCyc could be used to perform function-based retrieval of DNA or protein sequences, such as to prepare data sets for studies of protein structure–function relationships. Microbial geneticists who work with bacteria other than *E.coli* will find EcoCyc useful as a point of reference for gene functions, as well as for similarities and differences in gene–product relationships.

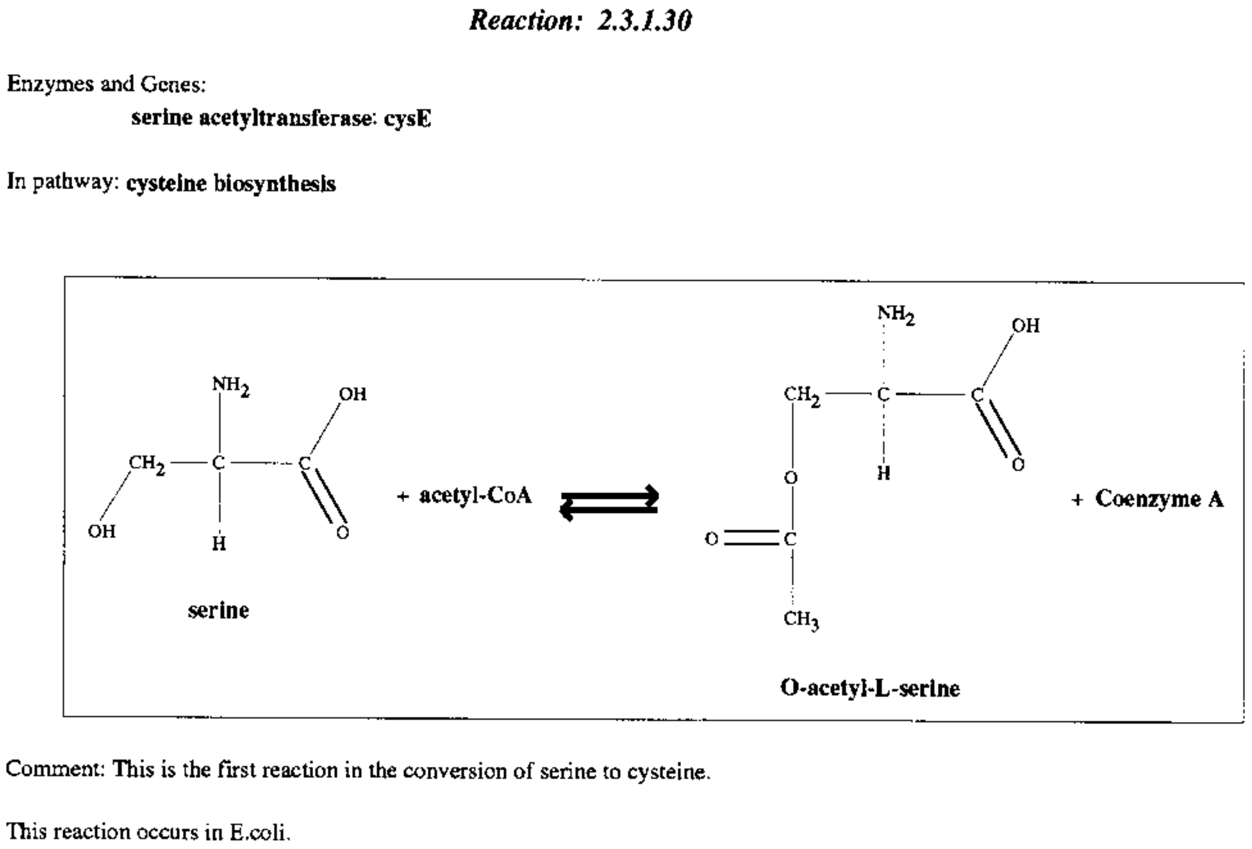
Studies of the metabolism

Scientists who study the evolution of the metabolism could use EcoCyc to search out examples of duplication and divergence of enzymes and pathways. Systematic computational studies of pathway evolution can compare related pathways from different organisms. EcoCyc provides a foundation for automatically generating simulations of the metabolism, although it lacks the kinetics data needed by most simulation techniques.

Pathway design for biotechnology

Biotechnologists seek to design novel biochemical pathways that produce useful chemical products (such as pharmaceuticals) or that catabolize unwanted chemicals such as toxins. EcoCyc provides the wiring diagram of *E.coli* K-12, which approximates the starting point for engineering. EcoCyc also describes the potential engineering variations that can result from importing *E.coli* enzymes into other organisms.

* To whom correspondence should be addressed



Reaction: 2.3.1.30

Enzymes and Genes:

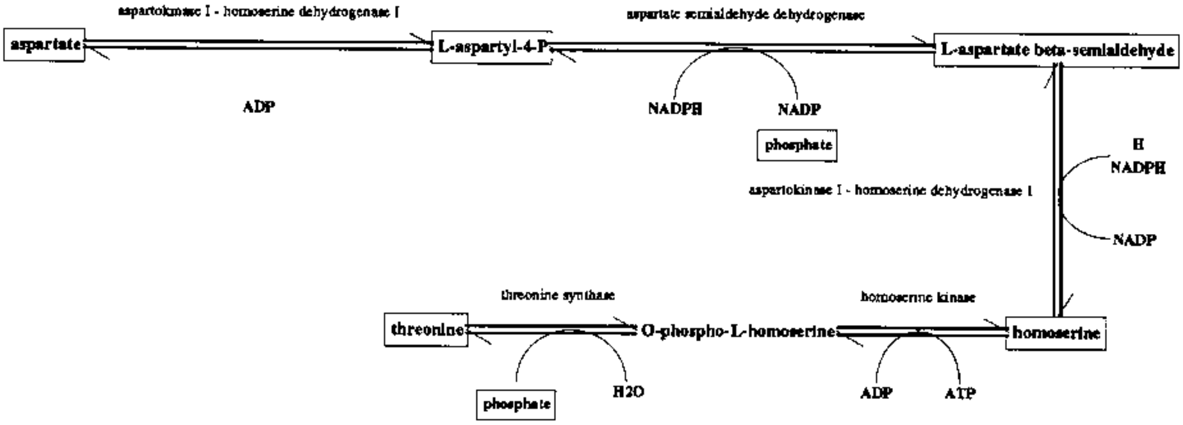
serine acetyltransferase: cysE

In pathway: **cysteine biosynthesis**

Comment: This is the first reaction in the conversion of serine to cysteine.

This reaction occurs in E.coli.

Figure 1. A reaction display window for the first reaction in the pathway for cysteine biosynthesis. The display shows properties of the reaction and lists related objects, such as the enzyme that catalyzes the reaction and the (one or more) genes that code for the enzyme.



aspartokinase I - homoserine dehydrogenase I
aspartate
L-aspartyl-4-P
ADP
aspartate semialdehyde dehydrogenase
L-aspartate beta-semialdehyde
NADPH
NADP
phosphate
aspartokinase I - homoserine dehydrogenase I
H
NADPH
NADP
threonine synthase
homoserine kinase
threonine
O-phospho-L-homoserine
homoserine
phosphate
H2O
ADP
ATP

Figure 2. The biosynthetic pathway for threonine.

THE ECOCYC GRAPHICAL USER INTERFACE

The EcoCyc GUI provides graphical tools for visualizing and navigating through an integrated collection of metabolic and genomic information (its retrieval capabilities are described in Retieval operations). For each type of biological object in the EcoCyc database the GUI provides a corresponding visualization tool. These tools dynamically query the underlying database to produce display windows such as are shown in Figures 1–3. Other displays are provided for genes, enzymes, compounds and for browsing genomic maps. All the 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 Karp and Paley (2).

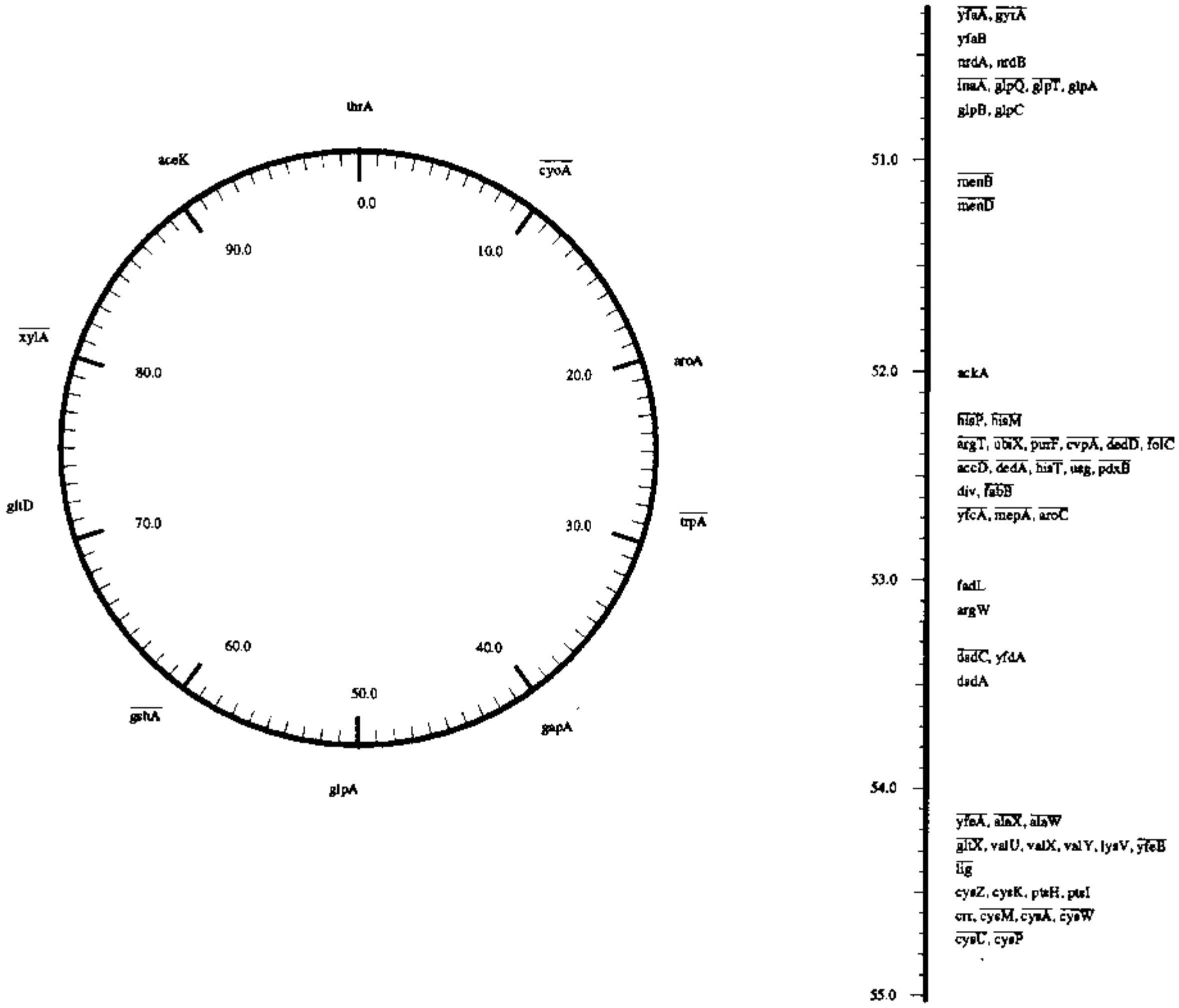


Figure 3. Circular map browser for the *E.coli* chrmosome.The bar at right shows a magnification of the region between 51 and 55 centisomes. Bars over gene names indicate counterclockwise transcription.

THE ECOCYC DATA

Our description of the EcoCyc database is intentionally more detailed than typical publications of this sort, because published descriptions of databases usually omit important information. Consider an analogy to a bench experiment. A publication describing a bench experiment carefully describes the experimental methods that were employed to aid the reader in understanding how the reported results were derived. An understanding of the experimental methods allows the reader to consider systematic errors and biases that may influence the results of the bench experiment. Similarly, computer databases are produced through complex procedures that contain biases and can introduce errors. An understanding of those procedures allows the user of a database to assess its potential strengths and weaknesses. A detailed understanding of the database conceptualization (schema) is another prerequisite for making proper use of a database. Many database authors wrongly assume that their cryptic table and attribute names convey the meaning of relationships that are often subtle. In fact, many databases are so poorly documented that the semantics of their attributes are impossible to discern.

The EcoCyc data are stored within a frame knowledge representation system (FRS), which is similar to an object-oriented database. FRSs use an object-oriented data model and have several advantages over relational database management systems (3). FRSs organize information within classes, collections of objects that share similar properties and attributes. The EcoCyc schema is based on the class hierarchy shown in Figure 4. All the biological entities described in EcoCyc are instances of the classes in Figure 4. For example, each gene is represented as an instance of the class Genes and every known polypeptide is an instance of the class Polypeptides. The current size of each class is shown in Table 1.

Each EcoCyc frame contains slots describing attributes or properties of the biological object that the frame represents or encoding a relationship between that object and other objects. For example, the slots of a polypeptide frame encode the molecular weight of the polypeptide, the gene that encodes it and its cellular location.

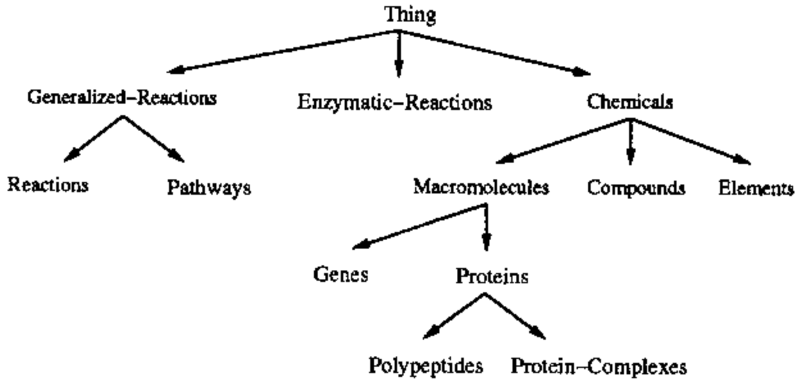


Figure 4. The top of the class hierarchy for the EcoCyc database. The arrows in this figure point from a general class of objects to a more specific class of objects. For example, we divide the class Proteins into the subclasses Polypeptides and Protein Complexes. The latter class is defined recursively, i.e. a component of a protein complex can itself be a protein complex.

Table 1. The number of objects in each EcoCyc class, both current and estimated total

	Current	Total (est.)
Reactions	580	650
Polypeptides	468	700
Pathways	89	100
Genes	2034	4000
Compounds	1181	?

The total column represents our best estimate as to the number of objects involved in *E.coli* intermediary metabolism.

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. In the future we plan to extend EcoCyc to describe various other aspects of cell function, including the preceding.

Genes

The majority of information on genes in EcoCyc was obtained from Rudd *et al.*'s EcoGene database version 6 (4). Rudd *et al.* have matched the restriction site patterns within sequenced genes to the restriction map, thereby determining the genomic map position of sequenced genes. Other fields we extracted from EcoGene were synonyms for the gene name, the direction of transcription and citations for each gene. Riley supplemented that information significantly by adding a list of unsequenced genes with known functions. She also classified all known genes into a comprehensive taxonomy of cellular function, which is included within EcoCyc (5,6). Riley connected most genes whose product is an enzyme to an EcoCyc object that represents that polypeptide and added a textual description of the product for genes whose product is not an EcoCyc object. EcoCyc contains 2034 genes, of which 1509 have genomic map positions. The genomic map can be viewed using both circular and linear map browsing tools that allow multiple levels of magnification within the chromosome.

Compounds

The class Chemicals subsumes all chemical compounds in the cell, such as macromolecules and smaller compounds that act as enzyme substrates, activators and inhibitors. It also includes the elements of the periodic table. This discussion focuses on small metabolites, which are instances of the class Compounds.

As a precursor project to EcoCyc Karp gathered information on chemical compounds from existing databases and from reference sources to form a database of metabolic compounds (7). He sought out data on compounds involved in the intermediary metabolism of *E.coli*, but data on some other compounds are also present, for example metabolites from other organisms. Among the properties encoded for compounds are synonyms, molecular weight, empirical formula, lists of bonds and atoms that encode chemical structures and two-dimensional display coordinates for each atom that permit the drawing of compound structures. EcoCyc contains 1181 compounds, of which 964 have recorded two-dimensional structures.

SRI International continues to update the compound data within EcoCyc, by adding new compounds, adding structures for existing compounds and correcting errors. Comprehensive compound data have been surprisingly useful to this project. Reaction equations in the literature use many different names to refer to the same compound. When we try to determine if two reactions are the same, we attempt to determine if their products and reactants are the same, making frequent use of our comprehensive compound synonym lists.

Reactions

The initial set of biochemical reactions in EcoCyc were derived from the ENZYME database (8), which Bairoch's group prepared by typing in the entries in Webb (9). We also downloaded the enzyme classification system of Webb (9) from ENZYME. Riley has added comments describing the metabolic role of many of these reactions. Because enzyme nomenclature concerns enzymes from all species, many of the reactions in the ENZYME database, and therefore in EcoCyc, do not actually occur in *E.coli*. Our project is to elucidate from the literature which reactions do occur in *E.coli*.

Riley also added new reactions to EcoCyc. For example, a number of the reactions catalyzed by *E.coli* have not been classified by the enzyme committee. In addition, some of the reactions in Webb (9) are written with a different specificity than the corresponding *E.coli* enzyme. This observation indicates a weakness of the enzyme nomenclature system. We cannot expect that a single reaction equation will accurately reflect the substrate specificities of enzymes from a variety of organisms.

Reaction frames contain information such as lists of reactants and products for the reaction equation, the EC number of the reaction and ΔG_0 for the reaction in the direction it is written. Reaction objects are linked to the pathway(s) that contains them and to the enzyme(s) that catalyzes them. EcoCyc contains 2901 reactions organized into the 258 classes defined by the enzyme committee. Of these, 580 reactions are known to occur in *E.coli*. Only 14 of the reactions have no EC number.

Proteins

Riley and Pellegrini-Toole have performed a comprehensive literature search for each enzyme, reaction and pathway. They used MEDLINE, Ungraham *et al.* (10) and biochemistry textbooks. They also followed citations in journal articles, leading to other pertinent papers. When the literature search was finished they used a standard text editor to enter information derived from the literature into a highly structured text file called a template file. Template files organize information as frames (such as enzymes and pathways) with labeled slots (attributes). The template files also permit association of chosen literature citations with the appropriate data. In each frame there are multiple opportunities for liberal comment, to describe the metabolic functions and the unique complex properties of the reaction or the enzyme. Among the topics covered by comments in EcoCyc are reaction mechanisms, subreactions of complex reactions, interactions of subunits of complex enzymes, formation of complexes with other proteins, breadth of substrate specificity, mode of action of inhibitors and activators, place and function of reactions in metabolic pathways, other reactions catalyzed by the protein and the relationship of the protein to other proteins catalyzing the same reaction.

Karp has developed a computer program that parses the template files to extract their constituent data items and then inserts those data items into the EcoCyc database. The parser program also performs consistency checks on the data and allows interactive correction of problems that are found. Consistency checkers can correct minor typographical errors and verify, for example, that the entry in a field that is supposed to contain a gene does in fact refer to a gene in the database. These tools have proven vital in detecting errors throughout the database.

In the EcoCyc schema all enzyme objects are instances of the class Proteins , which is partitioned into two subclasses: Protein Complexes and Polypeptides. These two classes have a number of common properties, such as molecular weight, pI, cellular location and a relationship to one or more catalyzed reactions. They differ in that Protein Complexes have slots that link them to their subunits, whereas Polypeptides have a slot that identifies their gene. We also record known sequence similarity relationships among a set of isozymes and we provide links to the SWISS-PROT and PDB entries for a polypeptide. Proteins are listed as a subclass of chemical compounds, since in some cases enzymes themselves are substrates in a reaction (such as phosphorylation reactions). The database contains 468 polypeptides and 238 protein complexes that comprise a total of 306 enzymes (i.e. 306 of the polypeptides and protein complexes have defined catalytic activities).

Figure 5 shows the EcoCyc display window for the enzyme serine acetyltransferase.

Enzymatic reactions

We define a high fidelity representation as a formal conceptualization (i.e. a portion of a schema) that allows a database to accurately capture subtleties of biology. For example, a number of other metabolic databases do not explicitly distinguish enzymes from reactions nor polypeptides from protein complexes. But the properties of a reaction (such as its ΔG_0 and its substrates) are independent of the enzyme(s) that catalyzes it and the properties of an enzyme (such as its molecular weight and amino acid sequence) are independent of the reactions it catalyzes. The relationships between enzymes and reactions are many to many, since one enzyme can catalyze many reactions and one reaction can be catalyzed by more than one enzyme. This distinction has led to interesting and perhaps counterintuitive observations. EC numbers are actually a property of reactions, rather than of enzymes, i.e. there is a one to one correspondence between reactions and EC numbers, but not between enzymes and EC numbers. An enzyme that catalyzes two reactions will have two EC numbers and two enzymes that catalyze the same reaction have the same EC number.

A further distinction is required because some properties of an enzyme are meaningful only in the context of a particular reaction that the enzyme catalyzes. Properties such as activators, inhibitors and cofactors pertain to the pairing of an enzyme and a reaction, because a single enzyme that catalyzes two reactions may be sensitive to different inhibitors for each reaction and we wish to capture this complex relationship. We capture it through a class called Enzymatic Reaction, which links an enzyme to a reaction that it catalyzes. Figure 6 depicts the objects and relationships that EcoCyc uses to portray the relationship between the reaction fumarate hydration and the enzymes that catalyze it (11).

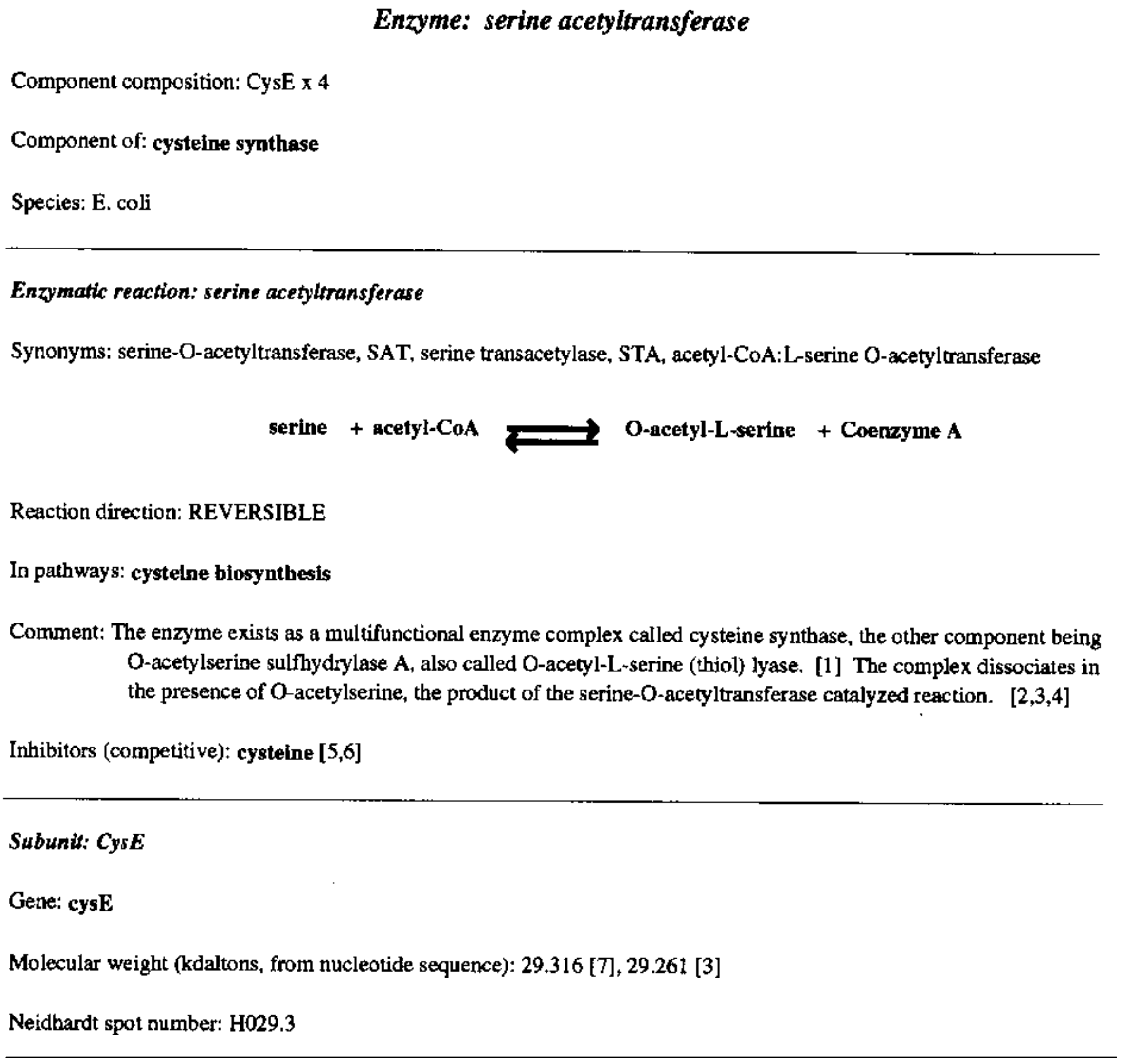
The slots of the Enzymatic Reaction class allow us to define four types of activators (competitive, allosteric, non-allosteric and those whose mechanism is not stated in the literature) and the analogous four types of inhibitors. Additional slots encode the cofactors, coenzymes and prosthetic groups of an enzyme.

The problem of encoding the substrate specificity of an enzyme is challenging. The EcoCyc schema provides three different means of encoding substrate specificity. Each approach has different advantages in terms of succinctness and in its ability to represent incomplete knowledge.

The first approach is to simply link an enzyme to a number of different reaction frames, each of which gives a precise reaction equation for a different combination of substrates. Secondly, we can write a generalized reaction equation that names classes of compounds, rather than exact compounds. For example, the equation

H_2O + an orthophosphoric monoester = phosphate + an alcohol

refers to a class of compounds called orthophosphoric monoester. This approach is much more succinct than the first approach, since we need not write out an equation for every variation of the reaction. However, the second approach may lead to over-generalization: Does the enzyme really accept any orthophosphoric monoester as a substrate or are only a limited set of them acceptable? We can explicitly list compounds that are members of the class of orthophosphoric monoesters to reduce the chance of over-generalization. Another approach would be to define the class by specifying a chemical fragment that is matched against



Enzyme: serine acetyltransferase

Component composition: CysE x 4

Component of: **cysteine synthase**

Species: E. coli

Enzymatic reaction: serine acetyltransferase

Synonyms: serine-O-acetyltransferase, SAT, serine transacetylase, STA, acetyl-CoA:L-serine O-acetyltransferase

serine + acetyl-CoA $\rightleftharpoons$ **O-acetyl-L-serine + Coenzyme A**

Reaction direction: REVERSIBLE

In pathways: **cysteine biosynthesis**

Comment: The enzyme exists as a multifunctional enzyme complex called cysteine synthase, the other component being O-acetylserine sulfhydrylase A, also called O-acetyl-L-serine (thiol) lyase. [1] The complex dissociates in the presence of O-acetylserine, the product of the serine-O-acetyltransferase catalyzed reaction. [2,3,4]

Inhibitors (competitive): **cysteine** [5,6]

Subunit: CysE

Gene: **cysE**

Molecular weight (kdaltons, from nucleotide sequence): 29.316 [7], 29.261 [3]

Neidhardt spot number: H029.3

Figure 5. A display window for the enzyme serine acetyltransferase.

all known compounds using a substructure searching algorithm (which we have in fact implemented for EcoCyc).

The third means of encoding substrate specificity is through the Alternative Substrates slot in the Enzymatic Reaction frame. For example, in the reaction

glutamate + ATP = glutamate-5-phosphate + ADP

we can specify that cycloglutamate is known to serve as an alternative substrate for glutamate. An advantage of this approach over the preceding ones is that if the product compounds generated from a given alternative substrate have not been observed experimentally (and are not obvious in theory from the chemistry of the reaction) we are still able to record the partial information that we have, i.e. we are not forced to write a complete reaction equation.

Another slot allows us to encode alternative cofactors in a similar fashion, such as if Mn^{2+} can substitute for Mg^{2+}.

Pathways

Pathway frames list the reactions that make up a pathway and describe the ordering of those reactions within a pathway. Information about the ordering of reactions within a pathway is encoded using a predecessor list representation (12), which for each reaction in a pathway lists the reactions that precede it in the pathway. This representation allows us to capture complex pathway topologies, yet does not require entering information that is redundant with respect to existing reaction objects. We have developed algorithms for deriving a full description of the pathway from the predecessor list (12).

If a reaction can be potentially catalyzed by more than one enzyme, but only one enzyme is physiologically active in a particular pathway (such as oxidative succinate dehydrogenase and anaerobic fumarate reductase), we can encode this restriction in the pathway frame within a slot called Enzyme Use by listing each reaction and the enzyme(s) that catalyzes it.

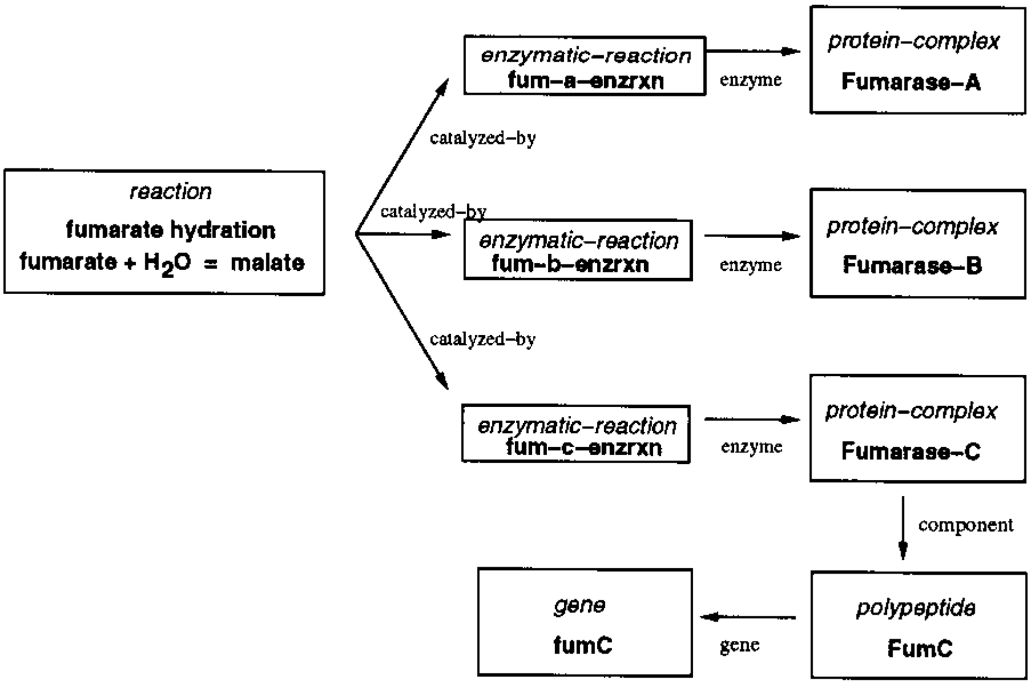


Figure 6. Relationships among EcoCyc frames. Boxes represent frames and arrows represent relationships between them. The fumarase reaction shown is catalyzed by three different isozymes in *E.coli*. Each isozyme is represented by a different object and each connects to the reaction object through a different enzymatic reaction. In addition, each isozyme is a homodimer; we show the connection of one of the fumarase C complexes to its component polypeptide and the connection of the polypeptide to the frame representing the gene that encodes it. Note that the database contains many relationships that are not shown in this figure. For example, the database includes the inverse of every relationship shown here.

The database uses objects called superpathways to define a new pathway as an interconnected cluster of smaller pathways. For example, a superpathway called 'complete aromatic amino acid biosynthesis' links together the individual pathways for biosynthesis of chorismate, tryptophan, tyrosine and phenylalanine. Superpathways are also defined using the predecessor list (12). EcoCyc currently contains 89 pathways and 11 superpathways.

The EcoCyc GUI uses automatic layout algorithms to generate drawings of linear, circular and tree structured pathways. The EcoCyc GUI allows the user to navigate from a pathway to a superpathway that contains it or vice versa. Pathway drawings can incorporate varying amounts of detail as specified by the user. Minimal detail shows only the names of compounds at branch points and on the exterior of the pathway; full detail shows all compound names, enzyme names and compound structures.

DATA VALIDATION

The EcoCyc data are subjected to several different validation checks to ensure their correctness. The database contains many consistency constraints that are automatically evaluated with respect to new entries, such as checking that the object listed as the product of a gene is in fact a polypeptide object or that the molecular weight of a protein is a positive number.

We also employ a reaction mass balancing program to search for database errors. The program evaluates all reactions for which every substrate of the reaction is a known compound within the EcoCyc database and the empirical formula of the compound is known. The program sums all atoms for each type of element for the products of the reaction and for the reactants and verifies that all atoms are conserved. (In fact, we allow hydrogen atoms to be non-conserved because of inconsistencies in ionization states across different compounds in the database.) This program has identified a number of errors in EcoCyc, including a dozen typographical errors in reactions obtained from the ENZYME database and errors in our compound structures. This program further illustrates the utility of including chemical compounds in a metabolic database.

Finally, Riley reviews each entry before its release. In the future we hope to enlist experts to review each pathway.

RETRIEVAL OPERATIONS

EcoCyc provides the user with two classes of database retrieval operations: direct retrieval through menus of predefined queries; indirect retrieval through hypertext navigation. For example, imagine that a user seeks information on the *hisA* gene, such as its map position and information about the enzyme it encodes. EcoCyc allows the user to call up an information window for that gene directly by querying the gene name.

The indirect approach consists of hypertext navigation among the information windows for related objects. Such navigation allows the user to find *hisA* by traversing many paths through the database. The user could issue a direct query to display the biosynthetic pathway for histidine and then click on the name of the enzyme at the last step in the pathway. The resulting information window for that enzyme will show the name of the gene (*hisA*) coding for the enzyme. Clicking on the gene name will display the information window for *hisA*. Alternatively, the user could query the compound histidine by name. The resulting window lists all reactions involving histidine. The user can then click on a reaction to navigate to its window, which lists all enzymes that catalyze the reaction plus all genes encoding those enzymes (including *hisA*).

Users invoke the queries using menus and dialog windows, rather than through a query language (we have partially implemented a declarative query language for EcoCyc). A distinctive aspect of EcoCyc is its extensive set of taxonomies. For example, EcoCyc includes the taxonomy of functions of gene products developed by Riley (5), a taxonomy of metabolic pathways and the taxonomy of reactions developed by the Enzyme Committee of the IUBMB (9). A user can query the gene taxonomy by first selecting a gene class from a menu of all classes (such as the class of genes coding for membrane proteins). Next, the user chooses one or more of the genes in that class from a second menu. The full set of queries supported by EcoCyc is as follows.

Gene queries

Get gene by name, Get gene by substring—Examples: Find *hisA*; Find all genes whose name includes 'his'

Get gene by class

Enzyme queries

Get enzyme by name, Get enzyme by substring

Get enzyme by pathway—Example: select from a menu of the enzymes in glycolysis

Reaction queries

Get reaction by pathway

Get reaction by EC number—Example: Find 1.2.3.4

Get reaction by class—select from a menu of all reactions in the EC class 1.2.3

Pathway queries

Get pathway by name, Get pathway by substring

Get pathway by class—Example: select from a menu of all pathways for amino acid biosynthesis

Compound queries

Get compound by name, Get compound by substring

Get compound by class

Get compound by substructure—Example: Find all compounds containing the substructure C-C-OH (substructures are specified using the SMILES language; 13)

Map queries

Create linear map display

Create circular map display

Zoom in on map; position is specified via mouse click, gene name or numerical map position

Add or remove genes from a partial map

When a query returns multiple answers the user can examine each answer in turn. The user can also employ a history list to return to a previous window.

SOFTWARE ARCHITECTURE

EcoCyc is implemented in Common Lisp using a graphical interface toolkit called the Common Lisp Interface Manager (CLIM). CLIM and Common Lisp are both highly portable, facilitating the delivery of EcoCyc on a variety of platforms. EcoCyc now runs on the Sun workstation under Common Lisp and CLIM products from Franz Inc. We plan a port to the PC in 1996.

EcoCyc builds on several software components. Metabolic pathway displays make use of the Grasper-CL graphing tool, developed at SRI International (14). Grasper-CL provides facilities for manipulation and display of graphs consisting of nodes and edges and provides a library of automatic layout algorithms. To store and manage the EcoCyc data we use an FRS called THEO developed at Carnegie-Mellon University and extended by our group at SRI International (15,16).

DISTRIBUTION

EcoCyc is available via the Internet in three forms.

A program for the Sun workstation bundles together the EcoCyc GUI and the EcoCyc database.

The EcoCyc database alone is available as a set of flatfiles.

The EcoCyc GUI is accessible on-line through the World Wide Web (WWW). The EcoCyc WWW pages describe all three types of access to EcoCyc. The WWW pages 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://www.ai.sri.com/ecocyc/ecocyc.html.

ACKNOWLEDGEMENTS

We thank Christos Ouzounis for adding links to SWISS-PROT and PDB to EcoCyc and for valuable discussions. 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 Karp,P. and Mavrovouniotis,M. (1994) *IEEE Expert*, **9** (2), 11–21.

2 Karp,P.and Paley,S. (1994) In Lim,H., Cantor,C. and Bobbins,R. (eds), *Proceedings of the Third International Conference on Bioinformatics and Genome Research.*

3 Karp,P. (1993) In Fortuner,R. (ed.), *Advanced Computer Methods for Systematic Biology: Artificial Intelligence, Database Systems, Computer Vision.*. The Johns Hopkins University Press, Baltimore, MD, p. 560.

4 Rudd,K.E., Bouffard,G. and Miller,W. (1992) In Davies,K.E. and Tilghman, S.M. (eds), *Genome Analysis*, Vol. 4, *Strategies for Physical Mapping*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 1–38.

5 Riley,M. (1993) *Microbiol. Rev.*, **57**, 862–952.

6 Riley,M. and Labedan,B. (1996) In Curtiss,R., Lin,E.C.C., Ingraham,J., Low,K.B., Magasanik,B., Neidhardt,F., Reznikoff,W., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella*, 2nd Edn. American Society for Microbiology, Washington, DC, in press.

7 Karp,P.D. (1992) *Comput. Appl. Biosci.*, **8** (4), 347–357.

8 Bairoch,A. (1994) *Nucleic Acids Res.*, **22**, 3626–3627.

9 Webb,E.C. (1992) *Enzyme Nomenclature, 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, New York, NY.

10 Ingraham,J., Low,K.B., Magasanik,B., Niedhart,F., Schaechter,M. and Umbarger,H.E. (eds) (1987) *Esherichia coli and Salmonella typhimurium*, Vols I and II. American Society of Microbiology, Washington, DC.

11 Karp,P. and Riley,M. (1993) In Hunter,L., Searls,D. and Shavlik,J. (eds), *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 207–215.

12 Karp,P. and Paley,S. (1994) In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 203–211.

13 Weiniger,D. (1988) *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.

14 Karp,P.D., Lowrance,J.D., Strat,T.M. and Wilkins,D.E. (1994) *LISP Symbolic Comput.*, **7**, 245–282. (See also *SRI Artificial Intelligence Center Technical Report* 521.)

15 Mitchell,T.M., Allen,J., Chalasani,P., Cheng,J., Etzioni,E., Ringuette,M. and Schlimmer,J.C. (1989) In *Architectures for Intelligence.* Erlbaum.

16 Karp,P.D. and Paley,S.M. (1995) In *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, pp. 751–758.