

# EcoCyc: Encyclopedia of *Escherichia coli* Genes and Metabolism

Peter D. Karp\*, Monica Riley<sup>1</sup>, Suzanne M. Paley, Alida Pellegrini-Toole<sup>1</sup> and Markus Krummenacker

Artificial Intelligence Center, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA and

<sup>1</sup>Marine Biological Laboratory, Woods Hole, MA 02543, USA

Received October 1, 1996; Accepted October 8, 1996

## ABSTRACT

The Encyclopedia of Genes and Metabolism (EcoCyc) is a database that combines information about the genome and the intermediary metabolism of *Escherichia coli*. It describes 2970 genes of *E.coli*, 547 enzymes encoded by these genes, 702 metabolic reactions that occur in *E.coli* and the organization of these reactions into 107 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 spans the space from sequence to function to allow scientists to investigate 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 silicio* model of *E.coli* metabolism that can be probed and analyzed through computational means.

## 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. It describes the 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. The EcoCyc graphical user interface (GUI) allows scientists to query, explore and visualize the EcoCyc DB. EcoCyc spans the space from sequence to function to allow scientists to investigate an unusually broad range of questions (5).

This article describes the scope of EcoCyc, 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. We request that users of EcoCyc cite this article in publications related to its use.

## MOTIVATIONS

EcoCyc can be viewed as an electronic review article because it is a carefully sifted collection of information drawn largely from (and containing 1400 citations to) the primary literature. EcoCyc is also designed to facilitate complex computations on genomic and metabolic data—to provide an *in silicio* model of *E.coli* that can be probed and analyzed through computational means. 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).

Because of its links to sequence DBs such as Swiss-Prot, EcoCyc could be used to perform function-based retrieval of DNA or protein sequences, such as to prepare datasets for studies of protein structure–function relationships. Scientists who study 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 performing simulations of the metabolism, although it currently lacks the kinetics data needed by most simulation techniques.

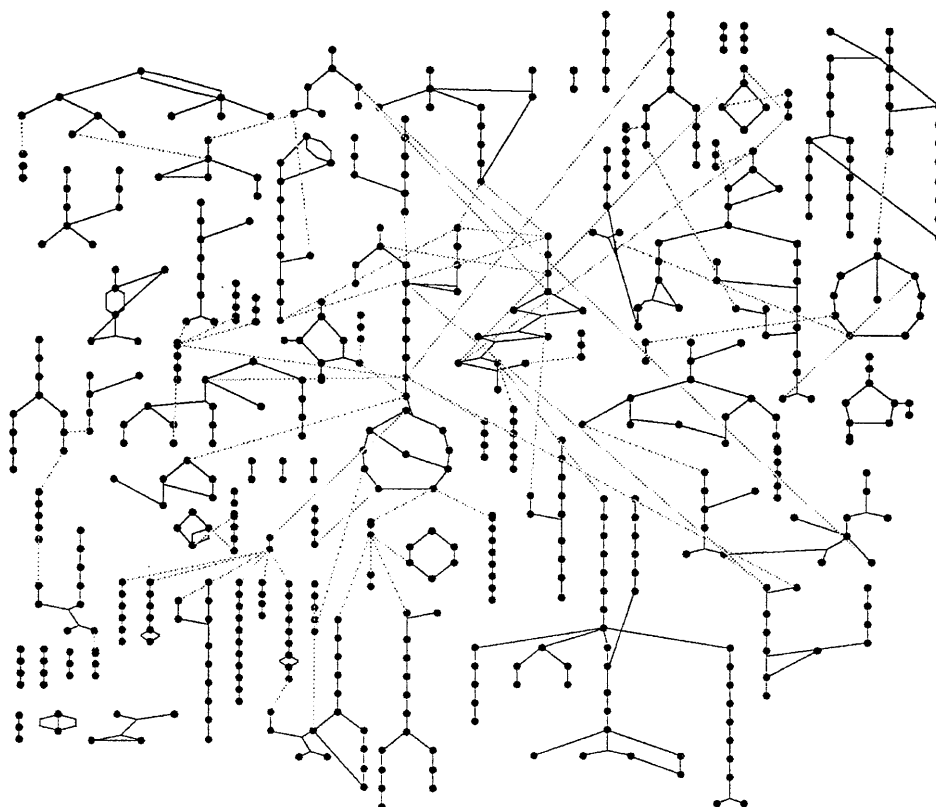
EcoCyc has been used to predict the metabolic complement of *Haemophilus influenzae* from its genomic sequence (14). That metabolic prediction was materialized in DB form and combined with the EcoCyc software to create an encyclopedia of the *H.influenzae* genome, called HinCyc. This metabolic-analysis technique extracts an added level of biological information from a genomic sequence, and provides a biological validation of the gene identifications predicted by sequence analysis.

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.

## RECENT ENHANCEMENTS

In the past year we have supplemented the EcoCyc data with 25 new pathways, and we finalized the descriptions of several pathways that previously were described only partially (these

\* To whom correspondence should be addressed. Tel: +1 415 859 6375; Fax: +1 415 859 3735; Email: pkarp@ai.sri.com



**Figure 1.** The overview shows a birds-eye view of the *E.coli* metabolic map. Each circle is a single substrate. Black lines are single reactions; gray lines connect two circles that represent the same chemical compound (although for aesthetic reasons, not all possible gray lines are drawn). The overview does not show reactions that have not been assigned to pathways.

pathways contained little information about their enzymes). We also added ~50 reactions of intermediary metabolism that play multiple metabolic roles, depending on conditions, and thus are not assigned to any particular pathway. Many citations were added to EcoCyc; it now contains 1400 literature citations. Gene information was downloaded from a new version (version 7) of the EcoGene database compiled by Rudd and Berlyn (2), and additional genes were added by our group. We added a second gene taxonomy based on product types, as described in the Genes section. Finally, we revised the representation of cofactors and of coenzymes within EcoCyc.

EcoCyc is now available on both Solaris and SunOS for the Sun workstation.

The GUI was enhanced to include two new visualizations: the Overview, and the Gene-Reaction Schematic. The Overview provides a birds-eye view of the entire metabolic map of *E.coli* as shown in Figure 1. Users can interrogate the Overview by moving the mouse over any reaction step or compound in the diagram, causing EcoCyc to display the name of the compound or reaction, and the name of its containing pathway, in a separate window. The user can also request that EcoCyc highlight one or more entities within the Overview, such as a specified compound, pathway, or enzyme. The Overview is also useful for comparative analysis of metabolic pathways, such as by highlighting the subset of pathways that are predicted to occur in *H.influenzae* (14). The Overview was generated through a combination of automatic layout of clusters of pathways and manual layout of the resulting clusters.

The Gene-Reaction Schematic provides a concise depiction of the relationships among genes, their polypeptide products, protein complexes that those polypeptides form, and reactions catalyzed by those proteins. These relationships exhibit a large range of complexity; the Schematic allows quick comprehension of complex relationships.

The structure of the EcoCyc class hierarchy is now visible in the GUI. The display windows for compounds, pathways, genes and reactions show the class that contains a given object. For example, this display of the reaction whose EC number is 1.3.5.3 shows that the class containing this object is EC 1.3.5. Clicking on that class name displays the class; that window lists all the instances of both the class (including 1.3.5.3) and its superclass (EC 1.3).

## THE ECOCYC GRAPHICAL USER INTERFACE

The EcoCyc GUI (4) provides graphical tools for visualizing and navigating through an integrated collection of metabolic and genomic information (its retrieval capabilities are described in Section 7). For each type of biological object in the EcoCyc DB, the GUI provides a corresponding visualization tool. These tools dynamically query the underlying DB to produce display windows such as shown in Figure 2, Figure 3 and Figure 4. Other displays are provided for genes, enzymes and compounds. All 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

*E. coli* Reaction: 2.7.1.40

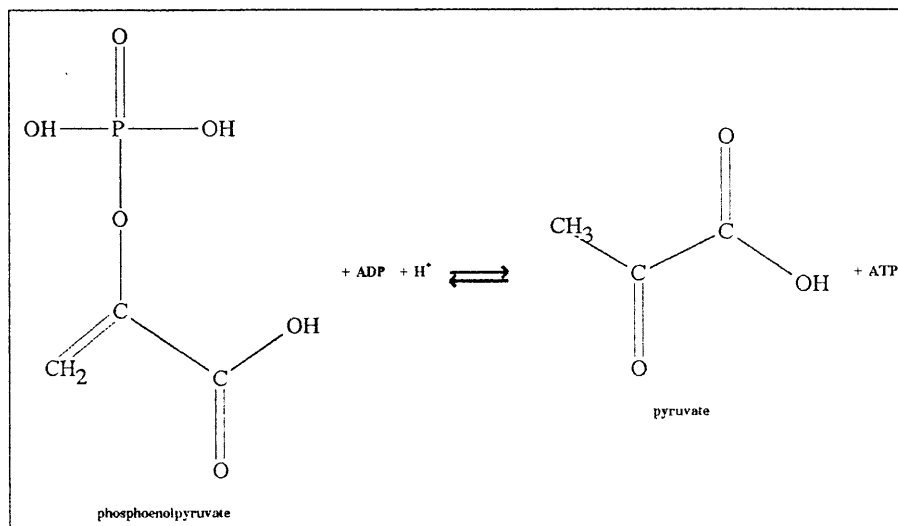
Superclass: 2.7.1 -- PHOSPHOTRANSFERASES WITH AN ALCOHOL GROUP AS ACCEPTOR

Enzymes and Genes:

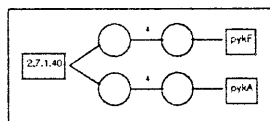
pyruvate kinase I: pykF

pyruvate kinase II: pykA

In pathway: fermentation, valine biosynthesis, glycolysis

 $\Delta G^{\circ}$  (kcal/mole): -7.5This reaction occurs in *E. coli*.

Gene-Reaction Schematic:



**Figure 2.** A reaction display window for the pyruvate-kinase reaction. 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. The gene-reaction schematic at the bottom shows that the reaction 2.7.1.40 is catalyzed by two different *E. coli* enzymes, each of which is a homotetramer.

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 (6).

## THE ECOCYC DATA

The EcoCyc data are stored within a frame knowledge representation system (FRS) called Ocelot. FRSs use an object-oriented data model, and have several advantages over relational DB 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 5 (10). All the biological entities described in EcoCyc are instances of the classes in Figure 5. For example, each *E. coli* 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. These statistics pertain to EcoCyc version 3.3. The next released version of EcoCyc, which should be available by the time this article is published, should be complete in that it will contain all known enzymes and pathways of *E. coli* small-molecule metabolism

(more enzymes will probably be discovered once the full *E. coli* sequence is known).

**Table 1.** The number of objects in each EcoCyc class

	Current
Reactions	702
Polypeptides	623
Pathways	107
Genes	2970
Compounds	1283

Each EcoCyc frame contains slots that describe attributes or properties of the biological object that the frame represents, or that encode a relationship among 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.

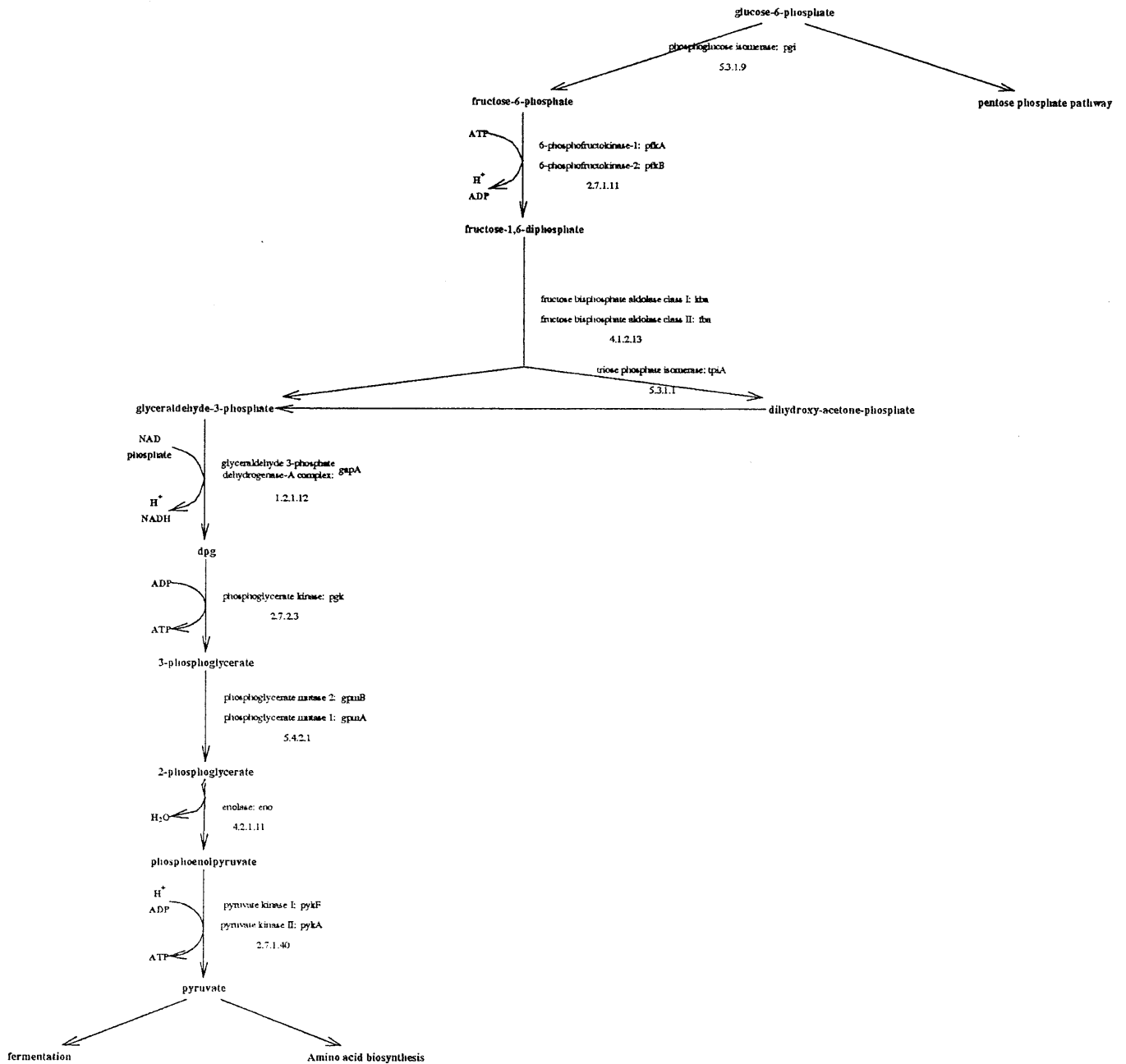


Figure 3. The glycolysis pathway.

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

**Genes**

Most information on *E.coli* genes in EcoCyc was obtained from the EcoGene DB version 7 (2). EcoGene provides synonyms for gene names, physical map positions for all sequenced genes, and

the direction of transcription for each gene. We supplemented the information in EcoGene significantly by adding descriptions of additional *E.coli* genes obtained from the literature and from SwissProt. EcoCyc contains 2970 genes, of which 2571 have assigned genomic map positions. The *E.coli* genomic map can be viewed with both circular and linear map-browsing tools that provide multiple levels of magnification within the chromosome.

EcoCyc classifies genes by using two classification systems. The first is based on the physiological role of the gene product (e.g., all genes whose products are involved in tryptophan biosynthesis are in a single category) (18). The second system is coarser, and assigns each gene to one of the following 10 product



compound. We determine if two reactions are the same by asking 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 DB (1), which Bairoch's group prepared by typing in the enzyme nomenclature system (20). We also downloaded the enzyme classification system (20) from ENZYME. We added comments describing the metabolic role of many of these reactions. Because the enzyme nomenclature concerns enzymes from all species, many of the reactions in the ENZYME DB do not actually occur in *E.coli*. EcoCyc reaction windows state whether or not we have evidence that a given reaction occurs in *E.coli*.

We have added more reactions to EcoCyc because a number of the reactions catalyzed by *E.coli* have not been classified by the enzyme committee. In addition, some of the reactions in (20) are written with 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 a family of enzymes from several organisms.

Reaction frames contain information such as lists of reactants and products for the reaction equation, the EC number of the reaction, and the  $\Delta G_0$  for the reaction in the direction it is written. Reaction objects are linked to the pathway(s) that contain them and to the enzyme(s) that catalyze them. EcoCyc contains 3038 total reactions organized into 269 classes defined by the enzyme committee; 702 of the reactions are known to occur in *E.coli* and 136 of the reactions have no EC number.

## Proteins

EcoCyc contains extensive information about *E.coli* enzymes and pathways that we obtained from the biomedical literature. We performed a comprehensive literature search for each *E.coli* enzyme, reaction, and pathway using Medline, the *E.coli-Salmonella* book (15) and biochemistry textbooks. We also carried out manual library searches for other pertinent papers by following citations in journal articles and in the *Science Citation Index*. Our original data entry method was to use 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 us to associate chosen literature citations with the appropriate data.

We developed a computer program that parses the template files to extract their constituent data items, and then inserts those data items into the EcoCyc DB. The parser program performs consistency checks on the data to correct minor typographical errors, and verifies, for example, that the entry in a field that is supposed to contain a gene does in fact refer to a gene in the DB. We recently discontinued use of the template files in favor of an interactive editing and browsing tool called the GKB Editor, which allows interactive entry of information directly into EcoCyc (16).

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 whether sequence-similarity relationships exist among a set of isozymes, and we provide links to the SwissProt and the PDB entries for a polypeptide. Proteins are listed as a subclass of chemicals since in some cases proteins themselves are substrates in a reaction (such as phosphorylation reactions). The DB contains 623 polypeptides and 319 protein complexes that comprise a total of 547 enzymes (i.e., 547 of the polypeptides and protein complexes have defined catalytic activities).

For each enzyme, we have written comments that address topics such as reaction mechanism, 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 relationship of the protein to other proteins catalyzing the same reaction.

## Enzymatic reactions

We define a high-fidelity representation as a formal conceptualization (that is, a portion of a schema) that allows a DB to accurately capture subtleties of biology (7,9). The design of the EcoCyc schema (class hierarchy) was motivated by several observations. The properties of a reaction (such as its  $\Delta G_0$  and its substrates) are independent of an enzyme(s) that catalyzes it, and an enzyme has a number of properties (such as molecular weight and amino-acid sequence) that are logically distinct from the reactions it catalyzes. The relationship between enzymes and reactions is many-to-many since one enzyme can catalyze several 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. That is, 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 the Enzymatic-Reaction, which links an enzyme to a reaction that it catalyzes, and essentially describes a single catalytic site within an enzyme.

The slots of the Enzymatic-Reaction class allow us to define four types of activators (competitive, allosteric, nonallosteric and those whose mechanism is not stated in the literature) and the analogous four types of inhibitors. By default, these activators and inhibitors are assumed to have been observed in enzymological studies; activators and inhibitors that are known to have physiological relevance are listed in the slot Physiologically-Relevant. Additional slots encode the cofactors, coenzymes and prosthetic groups of an enzyme. The EcoCyc schema also

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 (11).

## Pathways

Pathway frames list the reactions that make up a pathway, and describe the ordering of those reactions within the pathway. Information about the ordering of reactions within a pathway is encoded using a predecessor-list representation (7), 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 developed algorithms for deriving a full description of the pathway from the predecessor list (7).

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 the oxidative succinate dehydrogenase and the 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 catalyze(s) it.

The DB 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 (7). EcoCyc currently contains 107 pathways and 28 superpathways.

One possible point of confusion concerns the large discrepancy between the number of pathways found in EcoCyc, and in the *E.coli* subset of the EMP database (19). The number of pathways is larger in EMP because many of its pathways consist of a single reaction. In contrast, no EcoCyc pathways contain only one reaction. The average length of an EcoCyc pathway is six reactions.

The EcoCyc GUI uses automatic layout algorithms to generate drawings of linear, circular and tree-structured pathways. The 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 DB contains many consistency constraints that are automatically evaluated with respect to new entries, such as determining whether the object listed as the product of a gene is in fact an instance of the class Polypeptides, or that the molecular weight of a protein is a positive number.

We also employ a reaction mass-balancing program to search for DB errors. The program evaluates all reactions for which every substrate of the reaction is a known compound within the EcoCyc DB, 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 unconserved because of inconsistencies in ionization states across different compounds in the DB.) This program has identified a number of

errors in EcoCyc, including a dozen typographical errors in the reactions obtained from the ENZYME DB, and errors in our compound structures. This program further illustrates the utility of including chemical compounds in a metabolic DB.

Finally, we review each entry before its release. We have recently begun to enlist scientist experts to review each pathway.

## RETRIEVAL OPERATIONS

EcoCyc provides the user with two classes of DB retrieval operations: direct retrieval through menus of predefined queries, and 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 the *hisA* gene by traversing many paths through the DB. 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 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 queries using menus and dialog windows, rather than through a query language (although 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 two taxonomies of genes developed by Riley (18), a taxonomy of metabolic pathways, a taxonomy of chemical compounds, and the taxonomy of reactions developed by the enzyme committee (20). 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—Example: 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 (21)]

#### 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

#### Overview queries

– Highlight compounds or reaction steps using virtually all of the preceding types of queries

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 with 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.

EcoCyc builds on several software components. Metabolic pathway displays make use of the Grasper-CL graphing tool, developed at SRI (13). 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 Ocelot developed by our group at SRI; HyperTHEO is the predecessor of Ocelot and is described in (8). The World Wide Web (WWW) server capabilities are based on a software tool called CWEST (17).

## DISTRIBUTION

EcoCyc is available via the Internet in three forms:

- (i) A program for the Sun workstation (SunOS or Solaris) bundles together the EcoCyc GUI and the EcoCyc DB.
- (ii) The EcoCyc DB alone is available as a set of flat files.
- (iii) The EcoCyc GUI is accessible online through the WWW.

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://www.ai.sri.com/ecocyc/ecocyc.html>

## 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. (1994) The ENZYME databank. *Nucleic Acids Res.*, **22**, 3696–3627.
- 2 Berlyn,M.K.B., Brooks Low,K., Rudd,K.E. and Singer,M. Linkage map of *Escherichia coli* K-12, edition 9. In Neidhardt *et al.* (eds), *Escherichia coli and Salmonella, 2nd Ed.*, pp. 1715–1902.
- 3 Karp,P. (1993) Frame representation and relational data bases: Alternative information management technologies for systematics. In Fortuner,R. (ed.), *Advanced Computer Methods for Systematic Biology: Artificial Intelligence. Database Systems, Computer Vision*. The Johns Hopkins University Press, pp. 560.
- 4 Karp,P. (1996) *The EcoCyc Users Guide*, unpublished; see WWW URL <ftp://ftp.ai.sri.com/pub/papers/karp-ecocyc-guide.ps.Z>.
- 5 Karp,P. and Mavrovouniotis,M. (1994) Representing, analyzing, and synthesizing biochemical pathways. *IEEE Expert*, **9**, 11–21.
- 6 Karp,P. and Paley,S. (1994) Automated drawing of metabolic pathways. In Lim,H., Cantor,C. and Robbins,R. (eds), *Proceedings of the Third International Conference on Bioinformatics and Genome Research*. See also WWW URL <ftp://ftp.ai.sri.com/pub/papers/karp-bigr94.ps.Z>.
- 7 Karp,P. and Paley,S. (1994) Representations of metabolic knowledge: Pathways. 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.
- 8 Karp,P. and Paley,S. (1996) Integrated access to metabolic and genomic data. *J. Comp. Biol.*, **3**, 191–212.
- 9 Karp,P. and Riley,M. (1993) Representations of metabolic knowledge. 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.
- 10 Karp,P. and Riley,M. (1996) *Guide to the EcoCyc Schema*, unpublished. See WWW URL <ftp://ftp.ai.sri.com/pub/papers/karp-ecocyc-schema.ps>.
- 11 Karp,P., Riley,M., Paley,S. and Pellegrini-Toole,A. (1996) EcoCyc: electronic encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **24**, 32–40.
- 12 Karp,P.D. (1992) A knowledge base of the chemical compounds of intermediary metabolism. *Computer Appl. Biosci.*, **8**, 347–357.
- 13 Karp,P.D., Lowrance,J.D., Strat,T.M. and Wilkins,D.E. (1994) The Grasper-CL graph management system. *LISP and Symbolic Computation*, **7**, 245–282. See also SRI Artificial Intelligence Center Technical Report 521.
- 14 Karp,P.D., Ouzounis,C. and Paley,S.M. (1996) HinCyc: A knowledge base of the complete genome and metabolic pathways of *H. influenzae*. In States,D.J., Agarwal,P., Gaasterland,T., Hunter,L. and Smith,R. (eds), *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 116–124.
- 15 Neidhardt,F., III, Curtiss,R., Ingraham,J., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W., Riley,M., Schaechter,M. and Umberger,H.E. (eds) (1996) *Escherichia coli and Salmonella, 2nd Ed.* ASSI Press, Menlo Park, CA.
- 16 Paley,S. and Karp,P. GKB Editor user manual (1996). Available via WWW URL <http://www.ai.sri.com/~gkb/user-man.html>.
- 17 Paley,S.M. and Karp,P.D. (1996) Adapting EcoCyc for use on the world wide web. *Gene*, **172**, GC43–50.
- 18 Riley,M. (1993) Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.
- 19 Selkov,E., Basmanova,S., Gaasterland,T., Goryanin,I., Gretchkin,Y., Maltsev,N., Nenashev,V., Overbeek,R., Panyushkina,E., Pronevitch,L., Selkov,E., Jr and Yunus,I. (1996) The metabolic pathway collection from EMP: The metabolic pathways database. *Nucleic Acids Res.*, **24**, 26–29.
- 20 Webb,E.C. *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, 1992.
- 21 Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.