EcoCyc: The Resource and the Lessons Learned[‡]

Peter D. Karp, Monica Riley**

January 21, 1999

SRI International, 333 Ravenswood Avenue, Menlo Park CA 94025, USA, pkarp@ai.sri.com
**Marine Biological Laboratory, Woods Hole, MA 02543, mriley@mbl.edu

1 Introduction

The EcoCyc DB has several organizing principles. It is organized around the bacterium *E. coli* K–12. It is organized at the level of a review in that a given entry in the DB encodes information from a variety of sources about a single biological entity, such as an enzyme. A former organizing principle of the DB was to focus on information about enzymes and metabolic pathways; however, that focus is broadening to include transport, regulation, and other aspects of gene function.

EcoCyc is more than a DB — it is also a suite of software tools for visualizing and querying genomic and metabolic data. This chapter describes both the DB and the software tools. It surveys the content of the DB, and the mechanisms by which new data are acquired and validated. We close by discussing some of the lessons learned from the EcoCyc project.

2 The EcoCyc Data

The EcoCyc DB 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 software (GUI) allows scientists to query, explore, and visualize the EcoCyc DB. EcoCyc therefore integrates both genomic data and detailed descriptions of the functions of gene products. The EcoCyc data were drawn largely from (and contain 1650 citations to) the primary literature. In addition, some data were obtained from other DBs.

EcoCyc has potential uses in addition to its role as a reference source on *E. coli*. 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. EcoCyc provides a quantitative foundation for performing simulations of the metabolism, although it currently lacks the quantitative kinetics data needed by most simulation techniques.

EcoCyc has been used to predict the metabolic complements of *H. pylori* [5] and of *H. influenzae* from their genomic sequences [15]. The latter 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.

[‡]Appears in *Bioinformatics*, S. Letovsky, ed., Kluwer Academic Publishers, 1999.
[†]Portions of this publication were reprinted from *Nucleic Acids Research* 25:43 1997 by permission of Oxford University Press.

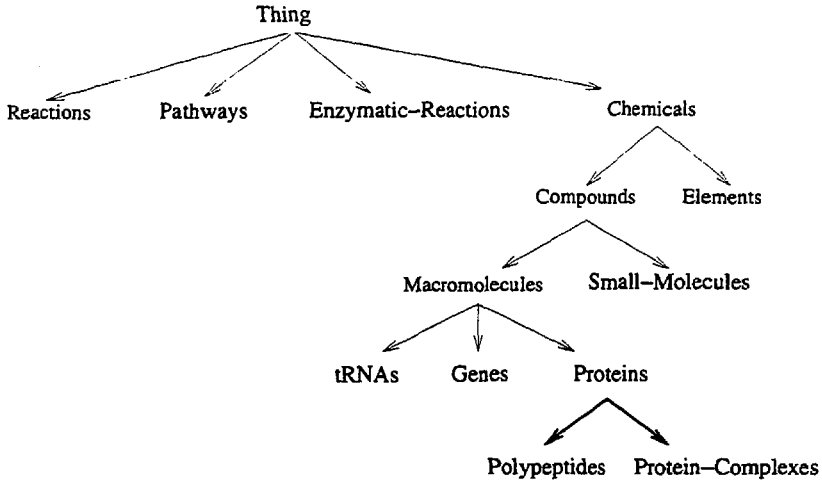


Figure 1: The top of the class hierarchy for the EcoCyc DB. The arrows in this figure point from a general class of objects to a more specific class of objects; for example, we divide the class `Proteins` into the subclasses `Polypeptides` and `Protein-Complexes`.

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.

3. The EcoCyc Graphical User Interface

The EcoCyc GUI [7] provides graphical tools for visualizing and navigating through an integrated collection of metabolic and genomic information (its retrieval capabilities are described in [13]). For each type of biological object in the EcoCyc DB, the GUI provides a corresponding visualization tool. There are tools for visualizing pathways, reactions, compounds, and so forth. These tools dynamically query the underlying DB for one or more objects and produce drawings specific to those objects. 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 side-compound names; they can also draw chemical structures for the compounds within a pathway [9].

4 Organization of the EcoCyc Data

The EcoCyc data are stored within a frame knowledge representation system (FRS) called Ocelot (its capabilities are similar to those of HyperTHEO, described in [10]). FRSs use an object-oriented data model, and have several advantages over relational DB management systems [6]. 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 1 [12]. All the biological entities described in EcoCyc are instances of the classes in Figure 1. 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`.

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

Table 1: The number of objects in several EcoCyc classes. The Enzymes row gives the number of polypeptides or protein complexes that catalyze a reaction. The numbers for Polypeptides and Protein Complexes also include some transport proteins.

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.

The scope of the data within EcoCyc is slowly expanding over time. The DB now describes most known *E. coli* genes; it describes those gene products that are enzymes involved in small-molecule metabolism, or that are tRNA synthetases, or that are involved in two-component signal transduction systems. It also describes gene products that are tRNAs. In the near future, we are planning to add descriptions of gene products that are transport proteins, or regulatory proteins.

New information is entered into EcoCyc using a combination of graphical editing tools. Some of these tools are specialized for entry of metabolic data, including graphical editors for reactions, compounds, and pathways. In addition, a domain-independent KB browsing and editing tool called the GKB Editor allows interactive editing of the EcoCyc class hierarchy, of a semantic-network view of the EcoCyc KB, and of individual EcoCyc frames; it also allows EcoCyc data to be transferred to a spreadsheet [17, 19].

Data validation techniques used in EcoCyc are described in [13].

We next describe the major classes of information within EcoCyc, the sources from which the information was obtained, and the visualization tools associated with those classes.

4.1 Genes

Most information on *E. coli* genes in EcoCyc was obtained from the EcoGene DB version 7 [2]. In the near future that information will be superseded by information from the full *E. coli* DNA sequence [3]. 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 3030 genes, of which 2571 have assigned genomic map positions. The map positions in EcoCyc version 3.7 were obtained from the EcoGene DB, but in the near future we wil obtain map positions from the full *E. coli* genomic sequence [3].

The visualization tool that generates gene-display windows lists information such as the map position of the gene on the *E. coli* chromosome (in units of centisomes, or hundredths of a chromosome), the class(es) to which the gene was assigned, and the direction of transcription. The gene product is listed (when known); when the product is an enzyme known to EcoCyc, the display shows the equation(s) of the reaction(s) catalyzed by the enzyme, and the pathways that contain those reactions.

We have classified EcoCyc genes according to two different 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, including enzymes and regulatory proteins, are in a single category) [20]. The second system is coarser, and assigns each gene product to one of the following classes: Enzymes, Regulators, Leaders, Membrane Proteins, Transporters, Structural Proteins, RNAs, Factors, Carriers, and products of unknown function.

4.1.1 The Gene–Reaction Schematic

The many-to-many relationships among genes, enzymes, and reactions can be complex. An enzyme composed of several subunits might catalyze more than one reaction, and a given reaction might be catalyzed by multiple enzymes. The *Gene–Reaction Schematic* depicts the relationships among a set of genes, enzymes, and reactions (see Figure 2). It is generated by starting with the object that is the focus of the current window (which is highlighted in the schematic), and then recursively traversing KB relationships from that object to related objects, such as from a gene to its product, or from a reaction to the enzyme(s) that catalyzes it. The schematic summarizes these complex relationships succinctly, and also constitutes a navigational aid — the user can click on an object in the schematic to cause EcoCyc to display that object.

The first schematic in Figure 2 means that the trpA gene encodes a polypeptide (the circle to the left of the box for the trpA gene) that forms a heterotetramer (the next circle to the left — the 2 indicates two copies) that also contains two copies of the product of the trpB gene. That complex in turn catalyzes reaction 4.2.1.20. The second schematic (reading down the column) depicts three isozymes (two homodimers and a homotetramer) that catalyze reaction 4.2.1.2. The third schematic depicts a bifunctional polypeptide, and the fourth schematic depicts a case where a homodimer of the TrpD polypeptide catalyzes one reaction, and a heterotetramer of TrpD and TrpE catalyzes a second reaction. The fourth schematic depicts two isozymes that each are heterotetramers. The fifth schematic depicts the ATP synthase protein, which consists of a large complex containing two subcomplexes.

Schematics also include modified forms of a protein (or tRNA) when relevant. For example, the schematic for the acyl carrier protein shows both a yellow circle for the unmodified form of the protein, and 13 orange circles, which represent different modified forms of the protein.

4.2 Reactions

The reactions within EcoCyc were gathered from biomedical literature on *E. coli*. In addition, we incorporated many non-*E. coli* reactions and 269 reaction classes that constitute the enzyme classification system [21] from the ENZYME DB [1]. EcoCyc therefore contains many reactions not found in *E. coli*, for reference purposes. EcoCyc reaction windows state whether or not we have evidence that a given reaction occurs in *E. coli*.

The reaction display window shows the class(es) containing the reaction within the classification of reactions. It shows the one or more enzymes that catalyze the reaction, the gene(s) that code for the enzymes, and the pathway that contains the reaction. The display shows the EC number for the reaction, and the reaction equation. Note that there exists a one-to-one mapping between EC numbers and reactions, but not between EC numbers and enzymes [11], therefore, we label reactions, and not enzymes, with the EC number. The standard change in Gibbs free energy of the reaction is listed when known.

4.3 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 by using Medline, the *E. coli–Salmonella* book [16], and biochemistry textbooks. We searched for other pertinent papers by following citations in journal articles and in the *Science Citation Index*.

In the EcoCyc schema, all enzyme objects are instances of the class of all proteins, which we call `Proteins`; it is partitioned into two subclasses: `Protein-Complexes` and `Polypeptides`. These two classes have several common properties, such as molecular weight (when the stoichiometry of the protein-complex is known), cellular location, and a link to any reactions catalyzed by the protein. They differ in that `Protein-Complexes` have slots that link them to their subunits, whereas `Polypeptides` have a slot that identifies their gene. We record whether sequence-similarity relationships exist among a set of isozymes, and we provide links to the SwissProt, PDB, and Swiss-Model 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).

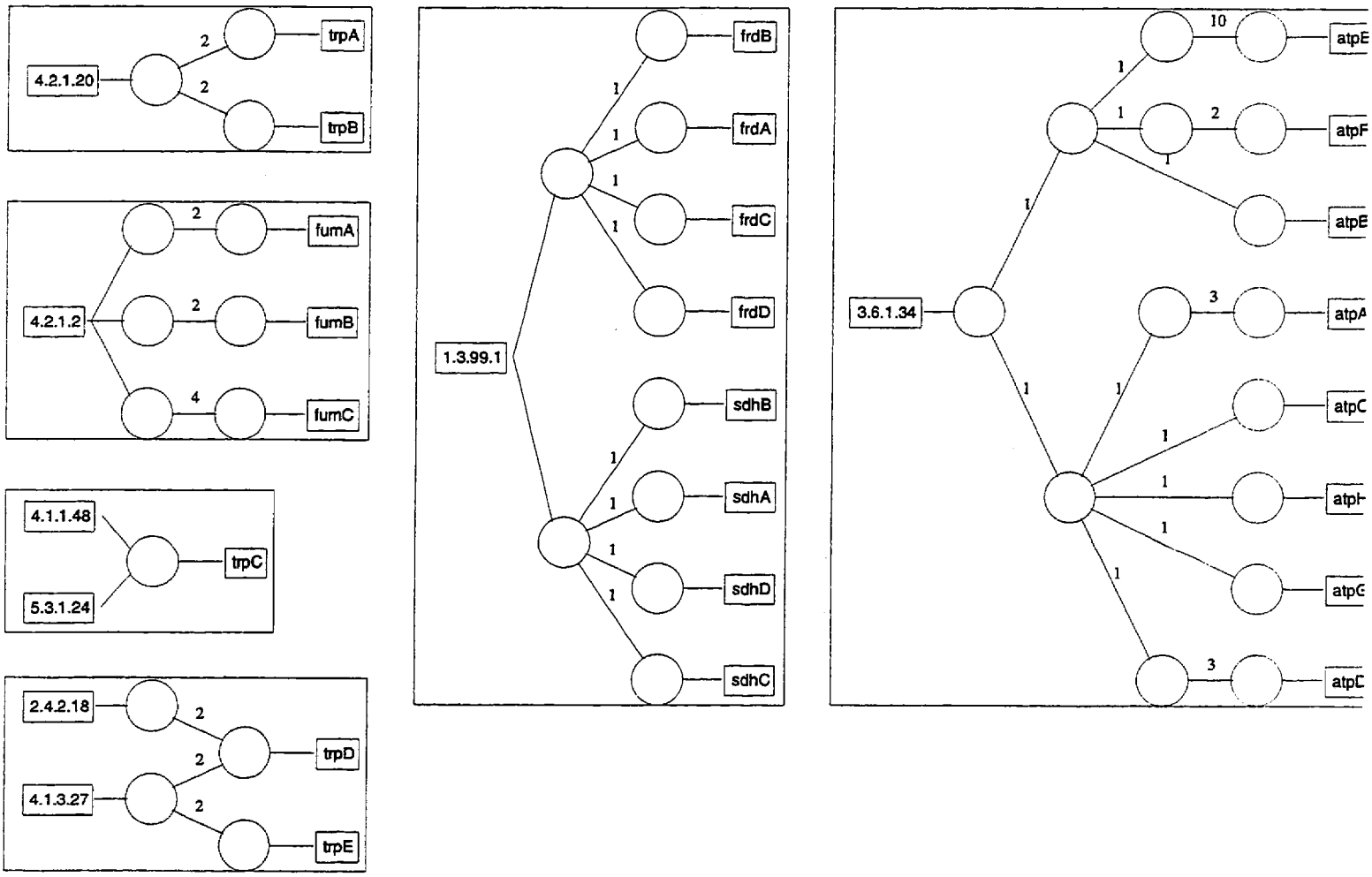


Figure 2: A set of gene–reaction schematics. Some of the reactions in the largest schematic have no assigned EC number. The boxes to the left represent reactions, the boxes on the right represent genes, and the circles in the middle represent proteins. The lines indicate relationships among these objects. The schematic is drawn in gene windows, reaction windows, and protein windows.

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.

Protein windows[1] are complicated because of the many-to-many relationship between enzymes and reactions (one enzyme can catalyze multiple reactions, and each catalytic activity of an enzyme can be influenced by different cofactors, activators, and inhibitors), and because many genes can encode the subunits of a protein complex. The protein window is potentially divided into sections to address these complexities.

The first section of the window lists general properties of the protein, such as synonyms and molecular weight. Subsequent sections of the window describe each catalytic activity of the protein, if it is an enzyme. Each activity section lists a reaction catalyzed by the enzyme, and the enzyme name (and synonyms) for that activity. The substrate specificity of the enzyme is described in some cases by listing alternative compounds that the enzyme will accept for a specified substrate. The cofactor(s) and prosthetic groups required by the enzyme are listed next,[2] along with any known alternative compounds for a specified cofactor. Activators and inhibitors of the enzyme are listed, qualified as to the mechanism of action, when known. In addition, this section indicates which of the listed activators and inhibitors are known to be of physiological relevance, as opposed to whether the effects are known purely because of *in vitro* studies. For a multifunctional enzyme, the descriptions of substrate specificity, cofactors, activators, and inhibitors are all tied to the enzyme activity to which they pertain. For more details on how this information is represented in EcoCyc, see [11].

4.4 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 [8], 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 graph description of the pathway from the predecessor list [8].

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 [8]. EcoCyc currently contains 123 pathways and 34 superpathways.

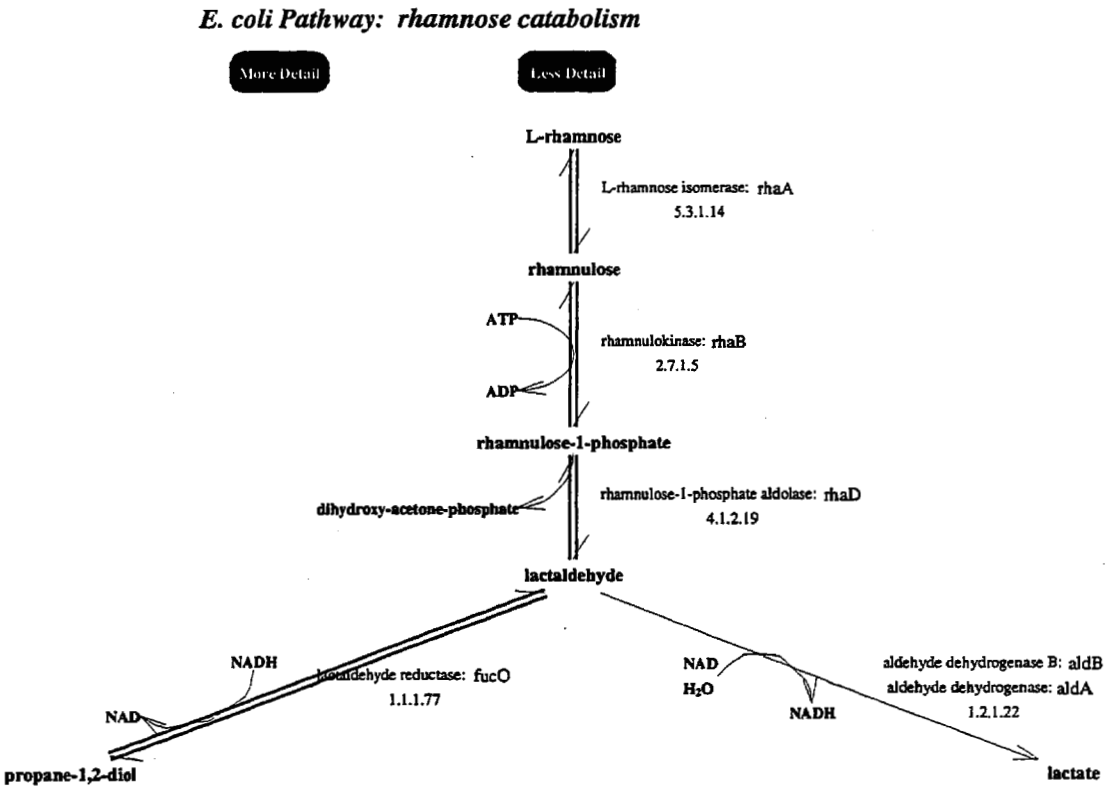
All pathway drawings in EcoCyc are computed automatically using pathway-layout algorithms (see Figure 3). EcoCyc can draw pathways at multiple levels of detail, ranging from a skeletal view of a pathway that depicts the compounds only at the periphery of the pathway and at internal branch points, to a detailed view that shows full structures for every compound, and EC numbers, enzyme names, and gene names at every reaction step. Users can select among these views by clicking on buttons labeled More Detail and Less Detail.

4.5 Compounds

The class `Chemicals` subsumes all chemical compounds in the *E. coli* cell, such as macromolecules and smaller compounds that act as enzyme substrates, activators, and inhibitors. It also includes some of the elements of the periodic table. Small metabolites contained in EcoCyc are reaction substrates, and enzyme cofactors, activators, and inhibitors.

EcoCyc contains 1294 compounds; two-dimensional structures are recorded for 965 of them. Among the properties encoded for compounds are synonyms for their names, molecular weight, empirical formula, lists

[1] See URL `http://ecocyc.PangeaSystems.com:1555//NEW-IMAGE?type=ENZYME&object=LACTALDREDUCT-CPLX` for an example.
[2] See [11] for a definition of these terms employed by EcoCyc.



Synonyms: rhamcat

Superclasses: **Carbon compounds**

Net reaction equation: rhamnose + ATP = glycerone phosphate + lactaldehyde + ADP

Superpathways: **fucose and rhamnose catabolism**

Locations of Mapped Genes:

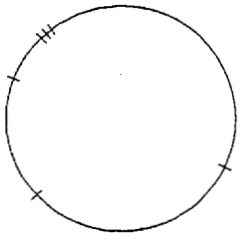


Figure 3: An EcoCyc drawing of the pathway for rhamnose catabolism. The small circle at the bottom of the pathway window depicts the positions on the *E. coli* genomic map of the genes that encode the enzymes within the current pathway. When the user moves the mouse over a given gene, its name is printed at the bottom of the EcoCyc window; clicking on the gene causes EcoCyc to display a window for that gene.

of bonds and atoms that encode chemical structures, and two-dimensional display coordinates for each atom that permit drawings of compound structures.

A compound display lists all EcoCyc reactions in which the compound appears, sorted by the pathways that contain each reaction. The display of chemical structures within compound windows uses a concept called superatoms, which is a hierarchical structuring of chemical structures. For example, the structure for succinyl-CoA is initially displayed with the word "CoA" in place of the structure of the CoA moiety. If the user clicks on the word CoA, however, the full structure of that moiety is displayed.

5 The Metabolic-Map Overview

The Overview diagram is a drawing of all known metabolic pathways of *E. coli*. In this diagram, each circle represents a single metabolite, and each line represents a single bioreaction. Neither the circles nor the lines

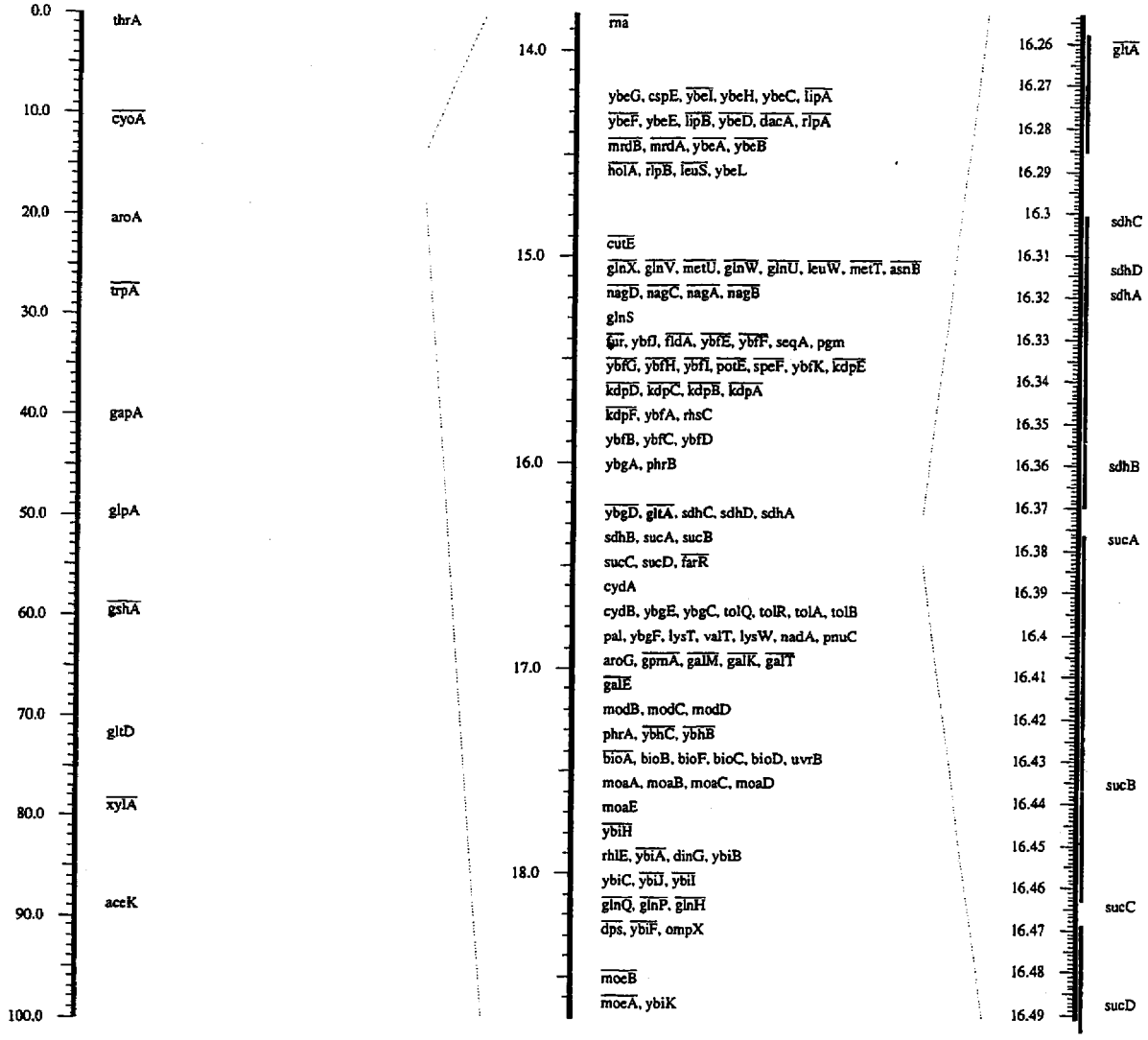


Figure 4: The EcoCyc linear map browser. Sections of the chromosome are shown at three resolutions. In the rightmost section, the coding region for each gene is shown with a vertical line.

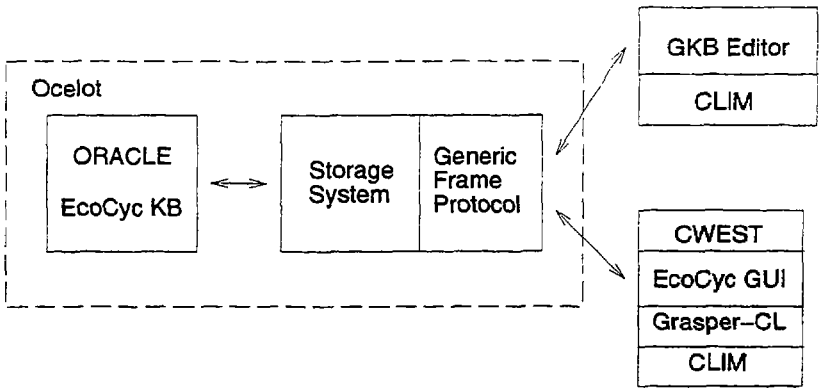


Figure 5: The software architecture of EcoCyc. The components include a graph-management system called Grasper-CL [14], an API for FBSs called the Generic Frame Protocol, and the CWEST tool for retrofitting

One lesson learned is that metabolic-pathway DBs are a useful addition to the repertoire of biological DBs. They provide a reference source on metabolic pathways. They have been successfully used to predict the metabolic complements of organisms from their genomic sequence [15]. Other potential applications of these DBs include pathway design for biotechnology [4], and simulation of metabolic pathways.

It is useful to organize pathway databases as a collection of multiple types of biological objects: pathways, reactions, enzymes, genes, and compounds. This organization is to be contrasted with that of biological databases such as GenBank, Swiss–Prot, and PDB, which contain only a single type of biological object.

Metabolic information is complex. The EcoCyc project has explored issues in the representation of metabolic pathways, reactions, and enzymes, and has developed an ontology for metabolic information that could be reused by other metabolic DBs [11, 8]. Most past biological DBs have encoded biological function within English text fields of the DB, whereas we have developed structured, declarative representations of function with which the user can query and compute. We have emphasized the development and publishing of our ontology because of the complexities of representing enzyme function and metabolic pathways. However, we believe that other biological-DB projects would benefit from a similar emphasis. Publication of ontologies both increases the understanding of the DB in the user community and aids the developers of other similar DBs.

We have designed graphical presentations of metabolic information for the EcoCyc GUI, and algorithms for generating those presentations, including automated layout algorithms for metabolic pathways [9].

The knowledge-acquisition problem has been a serious concern throughout the EcoCyc project. The problem is that of translating information in the scientific literature into a set of interconnected frames within a KB. The problem is particularly severe because EcoCyc contains so many different object classes. The process of describing a single complex metabolic pathway could involve creating several dozen different instances of four different classes, all of which must be properly linked. We have addressed this problem in three different ways: we have developed pathway-specific graphical tools for entry of pathway, reaction, and enzyme information; we also employ a general KB browsing and editing tool that is not independent of biology; finally, we have

References

[1] A. Bairoch. The ENZYME databank in 1995. *Nucl Acids Res*, 24:221–222, 1996.

[2] M.K.B. Berlyn, K. Brooks Low, K.E. Rudd, and M. Singer. Linkage map of *Escherichia coli* K–12, edition 9. In Neidhardt et al. [16], pages 1715–1902.

[3] F.R. Blattner, G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glasner, C.K. Rode, G.F. Mayhew, J. Gregor, N.W. Davis, H.A. Kirkpatrick, M.A. Goeden, D.J. Rose, B. Mau, and Y. Shao. The complete genome sequence of *Escherichia coli* k–12. *Science*, 277:1453–1462, 1997.

[4] C. Cameron and I. Tong. Cellular and metabolic engineering: An overview. *Applied Biochemistry and Biotechnology*, 38:105–140, 1993.

[5] J.-F. Tomb et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*, 388:539–547, 1997.

[6] P. Karp. Frame representation and relational data bases: Alternative information-management technologies for systematics. In R. Fortuner, editor, *Advanced Computer Methods for Systematic Biology: Artificial Intelligence, Database Systems, Computer Vision*, page 560. The Johns Hopkins University Press, 1993.

[17] S. Paley and P. Karp. GKB Editor user manual. Available via WWW URL `http://www.ai.sri.com/~gkb/user-man.html`, 1996.

[18] S.M. Paley and P.D. Karp. Adapting EcoCyc for use on the World Wide Web. *Gene*, 172(1):GC43–50, 1996.

[19] Suzanne M. Paley, John D. Lowrance, and Peter D. Karp. A generic knowledge-base browser and editor. In *Proceedings of the 1997 National Conference on Artificial Intelligence*, 1997.

[20] M. Riley. Functions of the gene products of *Escherichia coli*. *Microbiological Reviews*, 57:862–952, 1993.

[21] Edwin C. Webb. *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.