The EcoCyc Database

Peter D. Karp*, Monica Riley1, Milton Saier2, Ian T. Paulsen3, Julio Collado-Vides4, Suzanne M. Paley, Alida Pellegrini-Toole1, César Bonavides4 and Socorro Gama-Castro4

Bioinformatics Research Group, SRI International, 333 Ravenswood Avenue EK207, Menlo Park, CA 94025, USA, 1Marine Biological Laboratory, Woods Hole, MA 02543, USA, 2Department of Biology, University of California at San Diego, La Jolla, CA 92093-0116, USA, 3The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA and 4Programa de Biologia Molecular Computacional, Centro de Investigacion Sobre Fijacion de Nitrógeno, UNAM, Cuernavaca, A.P. 565-A, Morelos 62100, Mexico

Received September 18, 2001; Revised and Accepted October 17, 2001

ABSTRACT
EcoCyc is an organism-specific pathway/genome database that describes the metabolic and signal-transduction pathways of Escherichia coli, its enzymes, its transport proteins and its mechanisms of transcriptional control of gene expression. EcoCyc is queried using the Pathway Tools graphical user interface, which provides a wide variety of query operations and visualization tools. EcoCyc is available at http://ecocyc.org/.

INTRODUCTION
The EcoCyc database is a model-organism database for Escherichia coli K-12, and is a computational symbolic theory of the biochemical machinery of E.coli K-12 (1). As well as describing the genes and proteins of E.coli, EcoCyc goes beyond most model-organism databases because it provides structured symbolic descriptions of E.coli metabolic pathways, transport functions and gene regulation.

Intended uses of EcoCyc include the following. (i) EcoCyc is a resource for analysis of microbial genomes at the level of individual genes. Because the E.coli genome has a high fraction of genes whose functions were determined experimentally, it is an accurate reference for inferring gene function by sequence similarity. (ii) EcoCyc describes the subunit structures of many enzymes, and therefore could be used as training or validation datasets for algorithms that detect protein–protein interactions. (iii) Because it contains a symbolic description of the genetic network of E.coli, EcoCyc can serve as a test set for algorithms that infer genetic networks from gene-expression data. (iv) EcoCyc can be used for studies of pathway evolution. (v) EcoCyc is used as an aid in teaching biochemistry.

This article describes recent enhancements to EcoCyc, and how to access it. We request that EcoCyc users cite this article in publications related to its use. Version 5.6 of EcoCyc was released in June 2001.

PATHWAY TOOLS SOFTWARE AND DATABASE ENVIRONMENT
The Pathway Tools software that underlies EcoCyc provides query, editing and visualization operations for pathway/genome databases (2,3) (see http://bioinformatics.ai.sri.com/ptools/). The Pathway Tools software is an environment for functional bioinformatics—for managing, curating and computing with a functional genome annotation. The Pathway Tools utilize a frame knowledge representation system (FRS) called Ocelot (2,4). FRSs use an object-oriented data model that organizes information within classes: collections of objects that share similar properties and attributes.

THE EcoCyc DATA
Table 1 shows the current size of the principal EcoCyc classes.

Transcriptional regulation of gene expression
The current version 5.6 of EcoCyc includes information on regulation of transcription initiation, and organization of genes into transcription units (TUs). All data on gene regulation and TU organization were obtained from experimental characterizations of E.coli transcriptional regulatory mechanisms in the biomedical literature. A TU is a region of DNA that includes a single promoter, the transcription factor binding sites that modulate the rate of transcription initiation at that promoter, the genes that are transcribed from the promoter and the transcription terminator. A one-to-one relationship exists between TUs and promoters. A TU differs from an operon because by definition an operon must contain more than one gene, whereas a TU may contain one or more genes; an operon may also include several promoters and terminators, whereas a TU must contain a single promoter in this definition. Therefore, our approach defines a different TU for each promoter in an operon with multiple promoters.

The descriptions of genetic networks in EcoCyc were derived from RegulonDB version 3.2 (5), which was loaded into EcoCyc in 1999. Many updates, additions and corrections have been made since that time; curation of these data now occurs within EcoCyc.

EcoCyc currently contains descriptions of 100 DNA-binding transcriptional regulators [of the 314 total estimated regulators of E.coli (6)], 740 experimentally mapped promoters and 854 DNA-binding sites, as well as the clustering of 1185 genes into 629 TUs.

These data are incomplete in several respects. First, only 25% of all E.coli genes are clustered into TUs. Secondly, some of the defined TUs lack promoters because those promoters...
Table 1. The number of objects in version 5.6 of EcoCyc

<table>
<thead>
<tr>
<th>Object class</th>
<th>Object count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pathways</td>
<td>165</td>
</tr>
<tr>
<td>Reactions</td>
<td>2604</td>
</tr>
<tr>
<td>Enzymes</td>
<td>905</td>
</tr>
<tr>
<td>Transporters</td>
<td>162</td>
</tr>
<tr>
<td>Genes</td>
<td>4393</td>
</tr>
<tr>
<td>Transcription Units</td>
<td>629</td>
</tr>
<tr>
<td>Promoters</td>
<td>740</td>
</tr>
<tr>
<td>DNA-binding transcriptional regulators</td>
<td>100</td>
</tr>
<tr>
<td>DNA binding sites</td>
<td>854</td>
</tr>
<tr>
<td>Citations</td>
<td>3508</td>
</tr>
</tbody>
</table>

have not yet been physically mapped, although it is known that their genes are coregulated and produce a single transcript. Thirdly, some defined TUs do not include regulatory interactions because they await experimental determination, or their promoters are constitutive. For instance, the glnALG operon is described in EcoCyc by three different TUs, corresponding to the glnAp1, glnAp2 and glnL promoters. The glnAap2 sigma54 promoter and glnAp1 sigma70 promoter transcribe the three genes in this operon. The glnL internal promoter transcribes the last two genes.

SRI has developed several visualization tools to facilitate exploration of transcriptional regulatory data. EcoCyc gene-display windows now display a schematic diagram of the TU containing a gene, when known. Clicking on the TU produces a new TU window that lists information about each TU binding site, and the regulatory interactions of each TU binding site. The display window for a transcription factor displays all TUs that the transcription factor controls (its regulon). The overview diagram of the full metabolic map of E.coli contains an operation to highlight those metabolic-reaction steps and transport reactions controlled by a specified transcription factor.

Transport

The latest release of EcoCyc has considerably expanded coverage of membrane transport systems. Membrane transporters are responsible for uptake of metabolites and the export of metabolic end products, as well as being involved in other cellular processes. A total of 476 known and putative cytoplasmic membrane transport genes, corresponding to 304 probable transport systems, are currently represented in EcoCyc. A total of 177 distinct transport reactions are described in EcoCyc. In some cases single transporters mediate more than one transport reaction, and in other cases a single transport reaction may be mediated by more than one transporter.

Approximately 85% of the E.coli transporters have been annotated with the same detailed, literature-based approach that EcoCyc uses for E.coli enzymes and pathways. It is anticipated that curation of the remaining transporters will be complete by the next release of EcoCyc. Future objectives include expanding annotation targets to include outer membrane channels, protein secretion, all other aspects of membrane biology and membrane proteins of unknown function.

Transporters are represented in the EcoCyc database by one or more database objects that encode the transporter and its monomer subunits, if any. The transporter is linked to one or more (in the case of multifunctional transporters) objects that describe its function as a biochemical reaction. Our schema for reactions allows each substrate to be tagged with a cellular compartment, which if omitted defaults to the cytoplasm. These data are used to generate a cartoon diagram of each transporter that graphically represents the transport reactions. Transporters are now also displayed in the metabolic overview diagram with the directionality of transport indicated. By visual inspection, it becomes possible to probe the relationships between transported and metabolized compounds.

ADDITIONAL RECENT ENHANCEMENTS

The EcoCyc World Wide Web site at http://ecocyc.org/ has been expanded to also contain the MetaCyc database, and to contain pathway/genome databases for the following additional organisms, all created using the PathoLogic program (3):

- Bacillus subtilis
- Chlamydia trachomatis
- Haemophilus influenzae
- Helicobacter pylori
- Mycobacterium tuberculosis
- Mycoplasma pneumoniae
- Saccharomyces cerevisiae
- Treponema pallidum

The majority of proteins in EcoCyc now have World Wide Web links to the SWISS-PROT and PIR databases.

DISTRIBUTION

EcoCyc is available in four forms:

1. It is accessible online through the World Wide Web at http://ecocyc.org/ (this version supports a subset of the GUI functionality of the X-windows and PC versions).
2. An X-windows version of EcoCyc for the Sun workstation bundles together the Pathway/Genome Navigator software with the EcoCyc database.
3. A new PC version of EcoCyc bundles together the Pathway/Genome Navigator software with the EcoCyc database.
4. A flatfile version of EcoCyc is available for global analyses.

All four forms of access are free to academic institutions for research use; a fee applies to other forms of use. Contact ecocyc-info@ai.sri.com for information on obtaining the X-windows version, PC version or flatfile version. The EcoCyc World Wide Web site provides background information about the databases and software, and access to the publications produced by the EcoCyc project.

ACKNOWLEDGEMENTS

This work was supported by grant 1-R01-RR07861-01 from the Comparative Medicine Program at the NIH National Center for Research Resources. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

REFERENCES