D334–D337 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki108

EcoCyc: a comprehensive database resource for *Escherichia coli*

Ingrid M. Keseler, Julio Collado-Vides[1], Socorro Gama-Castro[1], John Ingraham[2], Suzanne Paley, Ian T. Paulsen[3], Martín Peralta-Gil[1] and Peter D. Karp*

SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, USA, [1]Program of Computational Genomics, CIFN, National Autonomous University of Mexico, Cuernavaca, A.P. 565-A, Morelos 62100, Mexico, [2]Section of Microbiology, University of California, Davis, One Shields Avenue, Davis, CA 95616, USA and [3]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Received September 15, 2004; Revised and Accepted October 18, 2004

ABSTRACT

The EcoCyc database (http://EcoCyc.org/) is a comprehensive source of information on the biology of the prototypical model organism *Escherichia coli* K12. The mission for EcoCyc is to contain both computable descriptions of, and detailed comments describing, all genes, proteins, pathways and molecular interactions in *E.coli*. Through ongoing manual curation, extensive information such as summary comments, regulatory information, literature citations and evidence types has been extracted from 8862 publications and added to Version 8.5 of the EcoCyc database. The EcoCyc database can be accessed through a World Wide Web interface, while the downloadable Pathway Tools software and data files enable computational exploration of the data and provide enhanced querying capabilities that web interfaces cannot support. For example, EcoCyc contains carefully curated information that can be used as training sets for bioinformatics prediction of entities such as promoters, operons, genetic networks, transcription factor binding sites, metabolic pathways, functionally related genes, protein complexes and protein–ligand interactions.

INTRODUCTION

Owing to its long history of being the focus of intense metabolic, biochemical and genetic investigations, *Escherichia coli* remains the best-studied bacterium and the primary reference organism for the exploration of function in other organisms. EcoCyc is a model organism database for *E.coli* that is a comprehensive reference resource as well as a tool for various types of computational exploration such as comparisons with genomes of other organisms. The Pathway Tools software, which supports EcoCyc, provides graphical display and editing capabilities as well as a convenient interface for querying and analyzing the database.

In recent years, coverage in EcoCyc has expanded from its original focus on metabolic pathways to include annotation and literature-based curation of all gene and protein functions, of enzymatic, transport and binding reactions, as well as transcriptional regulation, covering the entire genome. Development of EcoCyc is guided by a Scientific Advisory Board composed of outstanding scientists in various fields of *E.coli* research and model organism database development (see http://EcoCyc.org/advisors.shtml).

EcoCyc is a member of a larger collection of Pathway/Genome Databases (PGDBs) called BioCyc available at http://Biocyc.org/. We expect the number of BioCyc PGDBs to grow substantially in Fall 2004; please refer to the BioCyc website for details. Most members of the BioCyc collection model the genomes and pathways of a specific organism, and combine computationally predicted (1) and literature-based information in varying proportions. HumanCyc, for example, describes the metabolic map of *Homo sapiens*. The exception is MetaCyc, which describes experimentally elucidated metabolic pathways from more than 240 organisms (2) and is the reference database for predicting pathways in other BioCyc PGDBs.

THE EXPANDED SCOPE OF EcoCyc

A number of enhancements have been made to EcoCyc in the last three years (3). Our efforts to curate protein function based on the published literature have been expanded considerably. We are striving for complete coverage of all genes, proteins and RNAs in the *E.coli* genome by comments summarizing structural, functional and regulatory information from the

*To whom correspondence should be addressed. Tel: +1 650 859 4358; Fax: +1 650 859 3735; Email: pkarp@ai.sri.com

Table 1. Summary of data content in EcoCyc (Version 8.5)

Data type	Number
Genes	4497
Enzymes	1133
Protein comments	3461
Metabolic pathways	161
Metabolic reactions—small molecule metabolism	905
Signaling pathways	21
Transport reactions	202
Transcription units	956
Transcription start sites	1015
DNA-binding sites	1378
Transcriptional regulatory factors	148
Literature citations	8862

scientific literature. For example, the number of proteins covered by comments and functional annotations has more than tripled since 2002 to nearly 3500 in Version 8.5 of EcoCyc (see Table 1 and http://BioCyc.org/ecocyc/release-notes.shtml). Recent literature providing important insights into the properties of a gene or protein is incorporated into EcoCyc as quickly as possible, and literature references are in most cases provided as hyperlinks to the PubMed record.

An important recent addition to the information provided by EcoCyc is a set of icons allowing the user to evaluate the evidence underlying the curator's annotations (see below); these are provided in addition to hyperlinks to the pertinent primary literature. The icons indicate whether the information is based on experimental or computational evidence or based on human inference. Clicking on the icon displays a more detailed description of the evidence based on an evidence ontology (4), describing for example if a transcription start site was mapped by primer extension.

We are continuously incorporating new information, such as the recently released version of the *E.coli* K12 MG1655 genome sequence (GenBank accession number U00096, version U00096.2, GI:48994873; June 24, 2004), which corrects a significant number of sequencing errors present in the original genome sequence released in 1997 (5). In addition, extensive links to other databases such as Swiss-Prot and RefSeq have recently been added or updated. We also continue to curate the EcoCyc description of the *E.coli* metabolic network to reflect newly discovered pathways and enzymes. The EcoCyc network of small-molecule metabolism consists of 905 reactions catalyzed by 865 enzymes encoded by 961 genes.

Our curation procedure now includes partnering with outside experts on particular cellular systems to provide a more comprehensive literature overview and up-to-date coverage of the field. Special reviewers are acknowledged on the 'Credits' page (http://EcoCyc.org/contributors.shtml). Recently, this type of curation has been applied to the process of DNA repair. We have annotated both direct repair mechanisms, such as photolyase, as well as indirect repair mechanisms, such as nucleotide excision repair, base excision repair and homologous recombination. We have also curated 56 untranslated RNA species in EcoCyc.

Transport reactions in EcoCyc

The known and predicted membrane transporter complement of *E.coli* is now fully described within EcoCyc. A total of 202 transport reactions are described; all have been annotated with the same detailed, literature-based approach that EcoCyc uses for enzymes and pathways. Since the last publication describing the EcoCyc database (6), we have completed the curation of cytoplasmic membrane transporters and expanded coverage to include outer membrane channels, auxiliary transport proteins within transport systems, and protein secretion systems, such as the Sec and Tat pathways.

Transcriptional regulation in EcoCyc

Since merging with the RegulonDB (7) database in 1998, EcoCyc has incorporated extensive information on pathways that regulate the transcription initiation step of gene expression (3). EcoCyc's current contents on the elements supporting the regulatory network of *E.coli* are summarized in Table 1. The information on mechanisms of regulation gathered in both EcoCyc and RegulonDB is currently the largest known network of regulatory interactions of a bacterial cell, with 2393 specific interactions of transcription initiation (promoters and binding sites for regulators). The network includes more than 1000 mapped transcription initiation sites, which are regulated by nearly 1400 binding sites for specific transcriptional regulatory factors (TFs).

The names of TFs have been standardized in a manner that describes whether a TF acts as a repressor, activator or has a dual effect. Comments on regulatory proteins have been expanded and updated. Annotations now include, among others, the active conformation of TFs with the associated signal metabolites, the evolutionary family to which they belong, and whether they are autoregulated. The anatomy of regulatory regions upstream of operons and transcription units is also encoded in the database.

Most recently, we have expanded our scope to encompass regulation at all levels of gene expression and metabolism. We have initiated this effort by focusing on the conditions of anaerobiosis and utilization of carbon sources. This approach will produce a more integrated electronic description of the collection of regulatory mechanisms present in *E.coli*.

COMPUTING WITH THE EcoCyc DATA

The EcoCyc database, and other BioCyc databases, is accessible in several forms to facilitate computational exploration of the data, such as for machine learning or systems biology.

All BioCyc PGDBs are available for download in their entirety in several formats. These formats permit user parsing of the data by programs in languages such as Perl, and loading of the data into local database systems for processing. Here we list the formats available, and note that some formats include only a subset of the full PGDB contents because of limitations in the data model associated with that format. For complete documentation on these formats, see http://bioinformatics.ai.sri.com/ptools/flatfile-format.html. The formats are (i) column-delimited flat files (subset of PGDB contents), (ii) line-oriented attribute-value style flat files (subset of PGDB contents), (iii) SBML format (subset of PGDB contents), (iv) the BioPAX pathway-exchange format (http://www.biopax.org) and (v) Ocelot—a Lisp-based format.

For users who have downloaded and installed the Pathway Tools software, PGDBs can be accessed through an

Application Programming Interface (API) called GFP (Generic Frame Protocol). GFP supports programmatic interrogation and updating of PGDBs, and is the API through which Pathway Tools applications access PGDBs. GFP access is supported for the Java, Perl and Lisp languages. GFP API access is described in detail at http://bioinformatics.ai.sri.com/ptools/ptools-resources.html.

PATHWAY TOOLS SOFTWARE

Pathway Tools is more than just the software environment behind EcoCyc; it is a generic environment for building, editing, publishing and analyzing PGDBs. The software has evolved in many respects during the last three years to make new computational inferences, to support new datatypes and to include an expanded array of display and analysis tools. The software has also been ported to run on the Linux/Intel and Windows/Intel platforms in addition to SUN computers. The software runs both in a desktop application mode that allows users to create, edit and visualize PGDBs on their desktops, and in a Web mode that allows PGDBs to be queried and visualized through the Web.

PathoLogic—creation of new PGDBs through computational inference

Given an annotated genome, such as one in GenBank format, the PathoLogic component of Pathway Tools computationally generates a new PGDB containing the inferred metabolic pathways encoded in the genome. A program for inference of operons was recently added to PathoLogic (8); prediction of operons is based on a range of features including intragenic distance and functional relatedness of adjacent genes.

Metabolic pathways that are computationally predicted by PathoLogic often contain reactions for which no corresponding enzymes have been annotated within the genome. We call those reactions *pathway holes*. Pathway Tools now includes a *pathway hole filling* algorithm that predicts which genes within the genome may code for enzymes that fill these holes (9). When applied to a microbial genome, it can generate 50–100 new gene function predictions beyond those identified by classical genome annotation approaches.

Support for new datatypes

Pathway Tools has been extended to provide schema, display and editing support for the following additional datatypes: (i) Features of protein sequences such as metal-binding sites, disulfide bond locations, chemically modified residues, homology domains, repeats, signal sequences, DNA-binding regions and transmembrane regions. (ii) Introns, exons and alternative splice forms. (iii) Stereochemistry for chemical structures. (Many chemical compounds in our PGDBs now show stereochemical information in their structures. Editing of chemical structures is now supported via an interface to the JME chemical editor, written by Peter Ertl of Novartis.) (iv) An ontology of evidence codes which captures the evidence (computational, experimental, type of experiment, etc.) for the existence of various entities (pathways, protein functions, transcription units, etc.) in an organism (4). Pathway Tools has also been extended so that the pathway and operon predictors within PathoLogic decorate the pathway and operon PGDB objects that they create with evidence code information, indicating computationally predicted objects as such, and the editors within Pathway Tools have been extended to include functionality that allows users to interactively enter and modify evidence codes within PGDB objects.

Expanded display and analysis tools

Many enhancements have been made to the Cellular Overview diagram that provides a one-screen display of the full metabolic map of the cell. The diagram now includes additional cell structures such as transporters, the outer membrane with its proteins, and the periplasmic space with its proteins. Zooming of the diagram is now supported, and the layout of this diagram can be computed completely automatically by PathoLogic in a new PGDB, obviating the need for users to manually lay out the diagram. The Omics Viewer allows users to paint combined displays of gene expression, proteomics, metabolomics and reaction flux measurements on the diagram. For example, reaction flux measurements are painted as the colors of reaction lines in the diagram, and metabolomics measurements are painted as the colors of metabolite nodes in the diagram. Furthermore, these displays can be animated to show data from multiple time points or conditions.

Additional queries have been added to the desktop version of the software, such as queries to find RNA molecules, and to query proteins and small molecules according to multiple criteria. The main menu of the desktop software has been redesigned. The software also allows the user to retrieve arbitrary DNA sequences, in addition to the DNA or protein sequences associated with each gene.

A general import/export facility for PGDB objects has been added. Sets of objects can be exported to and imported from character-delimited files, which can be imported into a spreadsheet program, edited and re-imported. Objects can also be exported to an attribute-value format similar to MEDLINE format, and re-imported into a different database.

AVAILABILITY

In a new development, the EcoCyc data files are freely and openly available to all users, and the database can be redistributed. A binary executable version of Pathway Tools that includes EcoCyc and other BioCyc PGDBs is freely available to academic users and available for a fee to commercial users. The binary executable runs on SUN, Windows/Intel and Linux/Intel, and can run as both a desktop application and as an intranet web server. File and executable downloads are available via click-through license agreements at http://BioCyc.org/download.shtml, with new versions released four times a year.

ACKNOWLEDGEMENTS

We thank Martha Arnaud, Robert Gunsalus, Randy Gobbel, John Pick, César Bonavides-Martinez and Alberto Santos-Zavaleta for their contributions to EcoCyc and to Pathway Tools. This work was supported by grant RR07861 from the NIH National Center for Research Resources.

REFERENCES

1. Paley,S.M. and Karp,P.D. (2002) Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. *Bioinformatics*, **18**, 715–724.
2. Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*., **32**, D438–D442.
3. Karp,P.D., Arnaud,M., Collado-Vides,J., Ingraham,J., Paulsen,I.T. and Saier,M.H.,Jr (2004) The *E. coli* EcoCyc database: no longer just a metabolic pathway database. *ASM News*, **70**, 25–30.
4. Karp,P.D., Paley,S., Krieger,C.J. and Zhang,P. (2004) An evidence ontology for use in pathway/genome databases. In Altman,R.B., Dunker,A.K., Hunter,L., Jung,T. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing 2004*. World Scientific Press, Singapore, pp. 190–201.
5. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.
6. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc database. *Nucleic Acids Res*., **30**, 56–58.
7. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. and Collado-Vides,J. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res*., **32**, D303–D306.
8. Romero,P.R. and Karp,P.D. (2004) Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics*, **20**, 709–717.
9. Green,M.L. and Karp,P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76.