Chapter 6
The Multiple Scientific Disciplines Served by EcoCyc

Peter D. Karp

Contents

Abstract The EcoCyc database integrates information about the *E. coli* genome, its metabolic pathways, and its regulatory network. EcoCyc is in use by scientists from a variety of disciplines. Experimental biologists use it as a reference source on *E. coli*, and to leverage information about *E. coli* to the study of other microbes. Because the *E. coli* genome has the largest number of experimentally characterized genes of any organism, EcoCyc is used in the annotation of other microbial genomes by sequence similarity. EcoCyc has also been used in a number of global biological studies by computational biologists, and to provide training and validation datasets for the development of new bioinformatics algorithms. EcoCyc serves as a reference source for metabolic engineers, and it is used in microbiology education. The software behind EcoCyc, called Pathway Tools, has been used to develop EcoCyc-like databases for many other organisms. Pathway Tools provides powerful query and visualization capabilities, including tools to analyze high-throughput datasets by painting those datasets onto genome-scale diagrams of the metabolic network, the transcriptional regulatory network, and the complete genome map.

P.D. Karp (✉)
SRI International, 333 Ravenswood Ave AE 206, Menlo Park, CA 94025 USA
e-mail: pkarp@ai.sri.com

S.Y. Lee (ed.), *Systems Biology and Biotechnology of Escherichia coli*,
DOI 10.1007/978-1-4020-9394-4_6, © Springer Science+Business Media B.V. 2009

6.1 Introduction

The EcoCyc database has been under development since 1992 with the goal of serving several different scientific communities that require knowledge of the molecular parts of *E. coli*. EcoCyc has evolved from an initial focus on the metabolic pathways of *E. coli* to describe its complete genome and proteome, its metabolism and transport capabilities, and its regulatory network.

This chapter surveys the scientific disciplines served by EcoCyc. It discusses how these scientists use EcoCyc, and how the information and software tools provided by EcoCyc have been designed to serve their needs. The article also describes recently released software tools within EcoCyc, such as its Omics Viewers, the new graph tracks for visualization of ChIP-chip datasets, and the comparative analysis tools that support comparisons between any of the 370 organisms (including *E. coli*) that are supported within the BioCyc collection.

Our knowledge of who uses EcoCyc comes from a survey of EcoCyc users that we conducted in the spring of 2005, and from a citation analysis we performed for EcoCyc. To date, publications about EcoCyc (and the associated database RegulonDB (Gama-Castro et al. 2008, Salgado et al. 2004), which draws most of its content from EcoCyc) have been cited more than 500 times according to the ISI Web of Knowledge (http://www.isiwebofknowledge.com/). Scientists who use EcoCyc fall into the following groups:

- Experimental biologists who work with *E. coli*, other microbes, and higher organisms
- Computational biologists
- Bioinformaticists
- Metabolic engineers
- Educators

6.2 EcoCyc as a Reference for Experimental Biologists

EcoCyc is a knowledge resource for experimentalists who work with *E. coli*, other microbes, and higher organisms. Over the last 50 years a tremendous amount of information has been gathered on the genetics, biochemistry, and cell biology of *E. coli*, and continues to be amassed at a rapid pace. The pertinent literature is spread among a large number of scientific journals and many books.

Our 2005 Web-based survey asked what responders use EcoCyc for. Wet-lab biology usage indicated by the survey was as follows (each sentence contains responses from one survey question): study the biology of *E. coli* (30%); use *E. coli* as a model organism to study a particular aspect of biology (41%); use *E. coli* as a tool (e.g., for protein expression) (23%); other microbial research (31%); and other biological research (13%).[1] In the survey, 67% of responders said they use EcoCyc

[1] For questions in our survey that allow responders to select multiple choices, percentages refer to percent of responders who selected that answer, and do not add up to 100.

as a general *E. coli* reference tool; 19% use it as a tool for understanding other nonpathogenic bacterial species; 27% use it as a tool for understanding pathogenic bacterial species; and 28% use it for hypothesis generation (developing ideas for new experiments).

EcoCyc can be thought of as an online review article. In EcoCyc version 12.0, released in April 2008, 3,444 of the gene products described in EcoCyc (out of 4,472 total genes) contain mini-reviews authored by EcoCyc curators who summarize and cite the experimental literature for that gene product. The majority of these summaries are from 50 to 2,000 words in length. EcoCyc version 12.0 cites more than 16,000 peer-reviewed publications that have formed the basis for curation. Summaries are also found in other EcoCyc pages, including pathway and transcription unit pages. EcoCyc evidence codes describe the types of experimental evidence that support assigned gene functions. In version 12.0, functional assignments for 2,853 gene products are supported by experimental evidence, which is the highest in both relative and absolute terms of any model organism (Karp et al. 2007).

6.2.1 EcoCyc Analysis of Functional-Genomics Experiments

The use of DNA microarrays within the *E. coli* community has expanded tremendously. Proteomics and metabolomics work in *E. coli* is also increasing steadily. These “omics” methods yield large quantities of data that are difficult to analyze, but promise to produce new insights into cell function; 44% of our survey responders said they use EcoCyc for analysis of the *E. coli* regulatory network. EcoCyc facilitates analysis of functional-genomics data in two unique respects.

First, the extensive catalog of transcriptional regulatory circuits within EcoCyc puts known mechanisms of gene regulation at the fingertips of experimentalists, allowing them to focus on discovering new regulatory mechanisms rather than rediscovering known mechanisms. EcoCyc describes the regulation by 183 transcription factors of 1,492 promoters through regulatory interactions with 1,982 transcription factor binding sites. The majority of these regulatory interactions are based on experimental assays reported in the literature.

A new effort within the EcoCyc project aims to expand the types of cellular regulation encoded within EcoCyc. In 2007, the Pathway Tools software underlying EcoCyc was expanded to be able to capture, display, and edit six subtypes of regulation by attenuation, and curation of attenuation began. In 2008 we will be extending Pathway Tools to accommodate regulation by small RNAs, and translational regulation, and curation of these types of regulation will begin.

The second way in which EcoCyc facilitates analysis of functional-genomics data is via unique bioinformatics analysis capabilities, namely, three Omics Viewers that paint omics data onto global diagrams of *E. coli* cellular networks and of the *E. coli* genome. The same omics dataset can be viewed on all three diagrams so that it may be interpreted from different biological perspectives. Omics measurements are mapped to the same color scale on all three diagrams. Animation can be used on all three diagrams to display multiple measurements, which could reflect different time points, mutations, or treatments.

Examples of the three Omics Viewers are shown in Figs. 6.1, 6.2, and 6.3.

A new tool for analysis of ChIP-chip datasets is shown in Fig. 6.4. This tool, which we call graph tracks, is an extension of the genome-browser tracks capability. A ChIP-chip dataset is loaded into the Eco-Cyc genome browser (like the Omics Viewers, a data file can be uploaded via the EcoCyc Web site, or loaded into the desktop version of EcoCyc and Pathway Tools; the latter is recommended for frequent users because it runs faster and provides more capabilities). The dataset must be in GFF format (see http://www.sanger.ac.uk/Software/formats/GFF/). Data is plotted against the genome with intensity values depicted both as the Y coordinate and as color. Multiple graph tracks and normal horizontal tracks can be displayed simultaneously to compare multiple datasets.

6.2.2 Leveraging EcoCyc to the Study of Other Microbes

EcoCyc sits at the core of the BioCyc collection of Pathway/Genome Databases (PGDBs) for 379 organisms (Karp et al. 2005). For each of those organisms, BioCyc

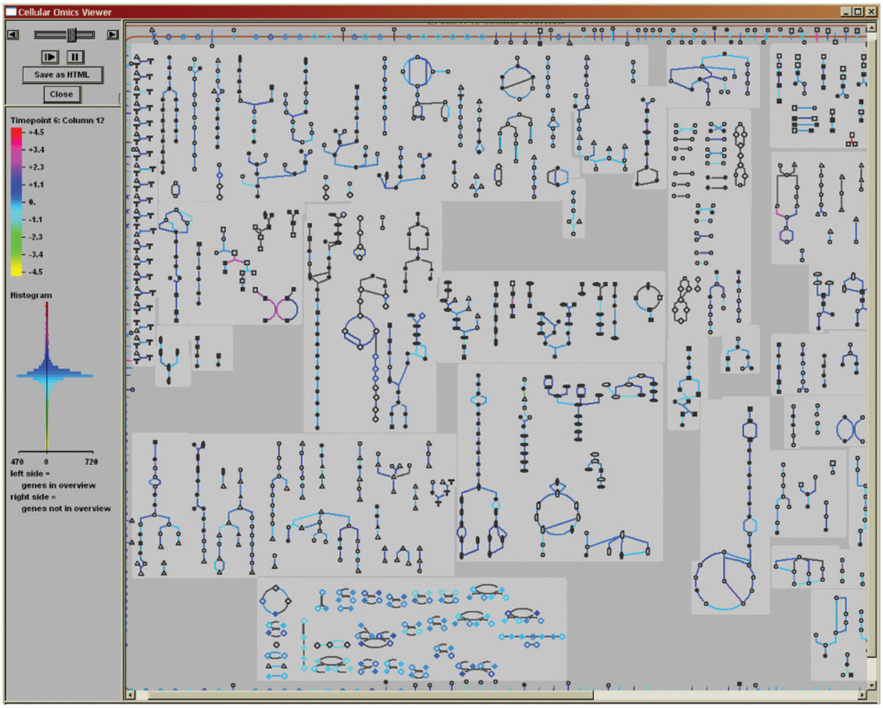


Fig. 6.1 The Cellular Omics Viewer. This image shows an *E. coli* gene expression dataset painted onto the *E. coli* metabolic network. The color assigned to each line (reaction) corresponds to the expression level of the gene coding for the enzyme that catalyzes that reaction. The controls in the upper left allow the user to stop and start animated displays within this diagram

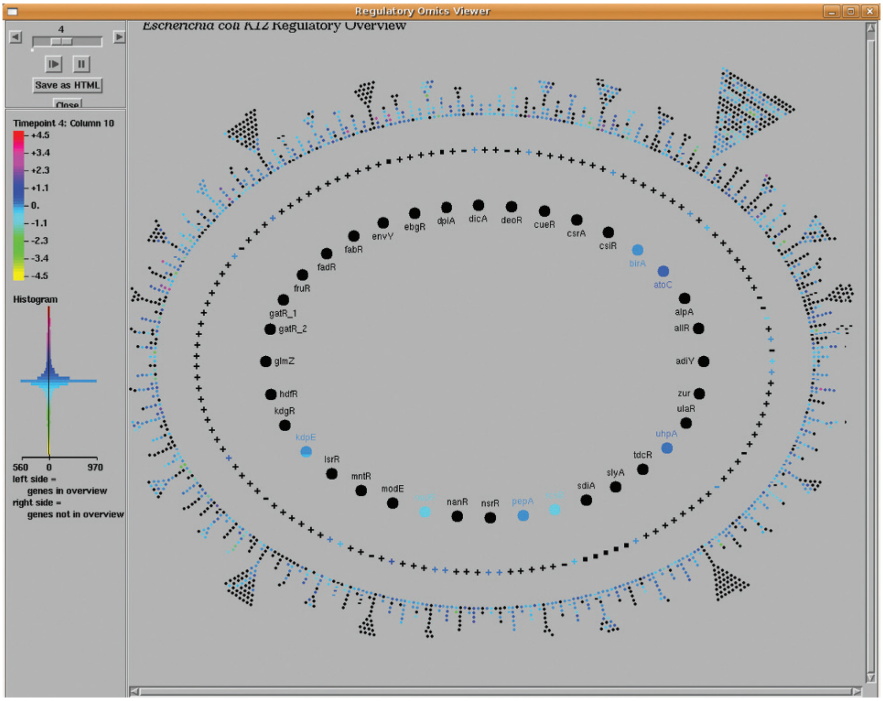
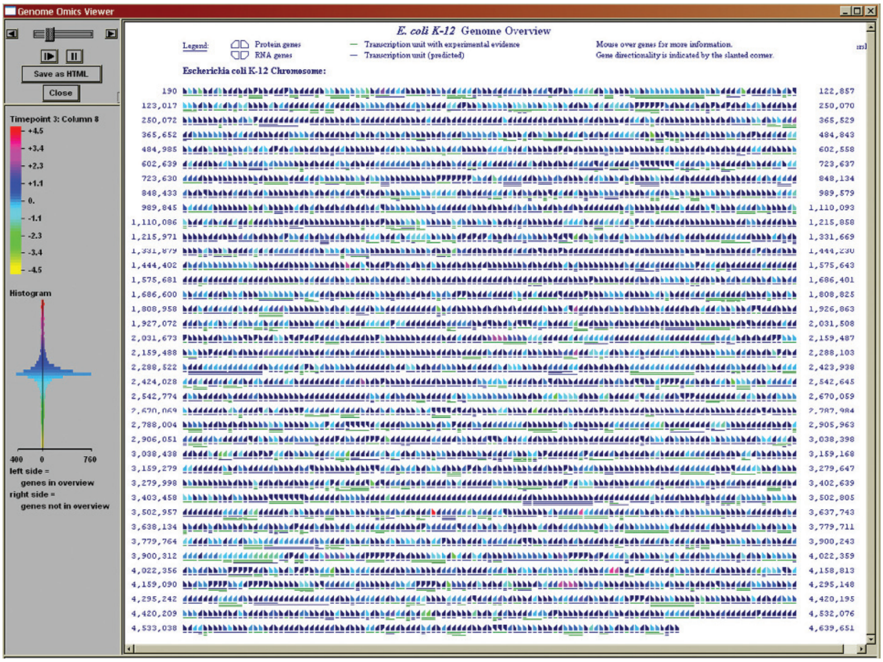


Fig. 6.2 The Regulatory Omics Viewer. This image shows an *E. coli* gene expression dataset painted onto the *E. coli* transcriptional regulatory network. The color assigned to each circle or square (genes) corresponds to the expression level of that gene. The innermost ring contains regulator genes (transcription factors and sigma factors) that have no regulatory inputs defined within EcoCyc. The middle ring contains regulator genes that do have defined regulatory inputs. The outer ring contains non-regulator genes. Genes in the outer ring are grouped into clusters such that two genes are assigned to the same cluster if those genes share the exact same set of regulators

contains their genomes, predicted metabolic pathways, and predicted pathway hole fillers (that is, genes that are predicted to code for enzymes missing from the metabolic pathways). For the bacteria, BioCyc also contains predicted operons. All the bioinformatics tools available for EcoCyc are also available for other organisms in BioCyc, including the genome browser and Omics Viewers.

The BioCyc.org site contains a powerful array of comparative genomics functionality that allows scientists who study other microbes to further leverage EcoCyc (46% of our users use EcoCyc to study other organisms besides *E. coli*), and also allows scientists who study *E. coli* to learn from its similarities to other organisms.

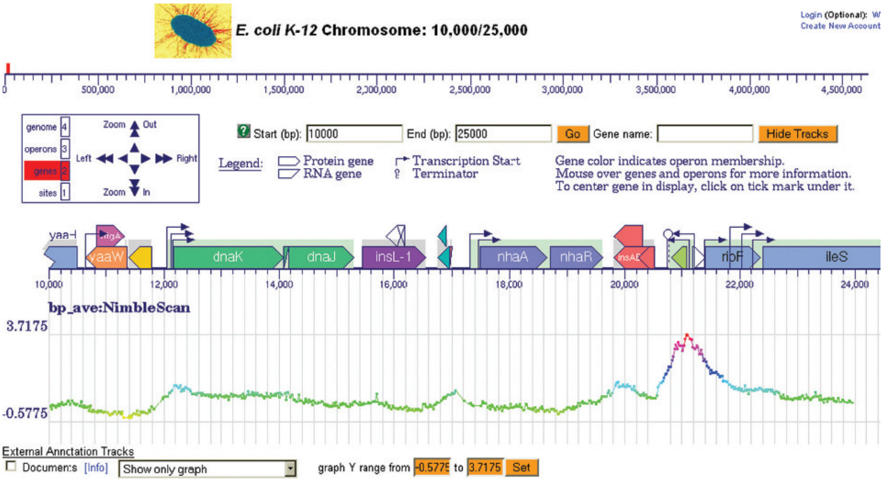
One comparative tool is the comparative genome browser. From a gene page in BioCyc (meaning, from a gene page for any organism in the BioCyc collection including EcoCyc), mid-way down the page is a button Align in Multi-Genome Browser. Clicking on the button will produce a list of all BioCyc organisms. Select the organisms of interest and click Submit. The resulting display will show



left side =
genes in overview
right side =
genes not in overview

Mouse over genes for more information.
Gene directionality is indicated by the slanted corner.

Fig. 6.3 The Genome Omics Viewer. This image shows the same *E. coli* gene expression dataset as shown in Fig. 6.1, painted onto the complete *E. coli* genome. Each “shark fin” represents a single gene. The color assigned to each gene corresponds to the expression level of the gene. Upward pointing genes code for proteins, downward pointing genes code for RNAs. The *left–right* directionality of each gene indicates its direction of transcription



Start (bp): 10000 End (bp): 25000 Gene name:

Gene color indicates operon membership.
Mouse over genes and operons for more information.
To center gene in display, click on tick mark under it.

□ Documents [Info] Show only graph

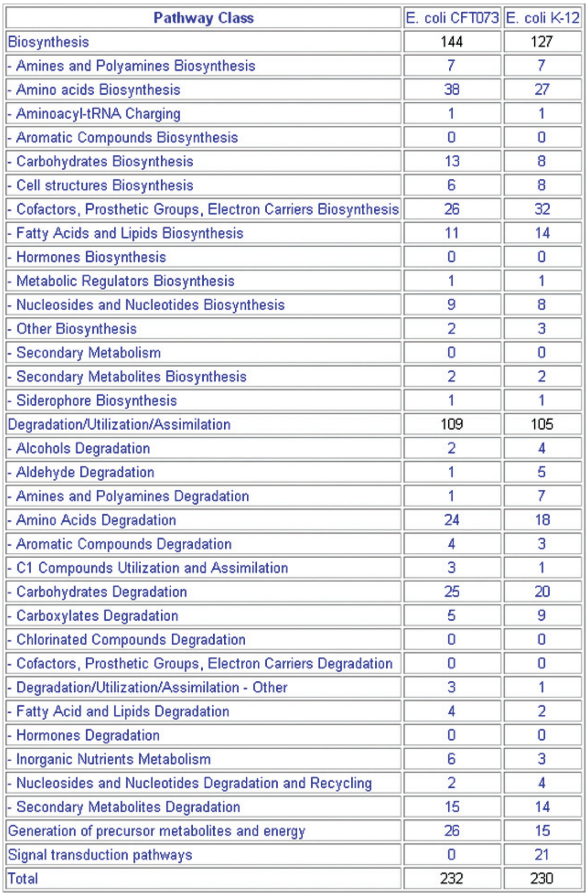
graph Y range from -0.5775 to 3.7175

Fig. 6.4 The EcoCyc genome browser with a graph track displayed. The graph track shows an X–Y plot of the intensity of RNA polymerase binding along the *E. coli* genome near the bottom of the figure. This ChIP-chip dataset was kindly provided by Dr. R. Landick of the University of Wisconsin

chromosomal regions for each organism, aligned at the orthologs of the starting genes. Genes drawn in the same color in this display are orthologs of one another (orthologs are defined as best bi-directional BLAST hits). The genome browser navigation controls can be used to zoom or translate the genome map display.

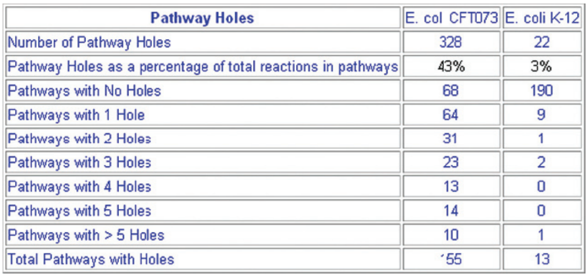
A second set of comparative tools is available from the Comparative Analysis link on the main BioCyc query page (http://biocyc.org/server.html). These tools will generate comparative reports across many dimensions of a PGDB. Several report types are available. One report compares the metabolic pathway complements of the selected organisms. Another report compares the metabolic reaction complements. Another compares transporter complements. Reports are also available to compare proteins, metabolites, and transcription units.

Each report contains several sections. For example, Figs. 6.5 and 6.6 show sections of the pathway report. The reports contain tables that contain summary statistics. To drill down to the data from which those summary statistics were integrated, click on a cell within the table. For example, clicking on a row name will produce



Pathway Class	E. coli CFT073	E. coli K-12
Biosynthesis	144	127
- Amines and Polyamines Biosynthesis	7	7
- Amino acids Biosynthesis	38	27
- Aminoacyl-tRNA Charging	1	1
- Aromatic Compounds Biosynthesis	0	0
- Carbohydrates Biosynthesis	13	8
- Cell structures Biosynthesis	6	8
- Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis	26	32
- Fatty Acids and Lipids Biosynthesis	11	14
- Hormones Biosynthesis	0	0
- Metabolic Regulators Biosynthesis	1	1
- Nucleosides and Nucleotides Biosynthesis	9	8
- Other Biosynthesis	2	3
- Secondary Metabolism	0	0
- Secondary Metabolites Biosynthesis	2	2
- Siderophore Biosynthesis	1	1
Degradation/Utilization/Assimilation	109	105
- Alcohols Degradation	2	4
- Aldehyde Degradation	1	5
- Amines and Polyamines Degradation	1	7
- Amino Acids Degradation	24	18
- Aromatic Compounds Degradation	4	3
- C1 Compounds Utilization and Assimilation	3	1
- Carbohydrates Degradation	25	20
- Carboxylates Degradation	5	9
- Chlorinated Compounds Degradation	0	0
- Cofactors, Prosthetic Groups, Electron Carriers Degradation	0	0
- Degradation/Utilization/Assimilation - Other	3	1
- Fatty Acid and Lipids Degradation	4	2
- Hormones Degradation	0	0
- Inorganic Nutrients Metabolism	6	3
- Nucleosides and Nucleotides Degradation and Recycling	2	4
- Secondary Metabolites Degradation	15	14
Generation of precursor metabolites and energy	26	15
Signal transduction pathways	0	21
Total	232	230

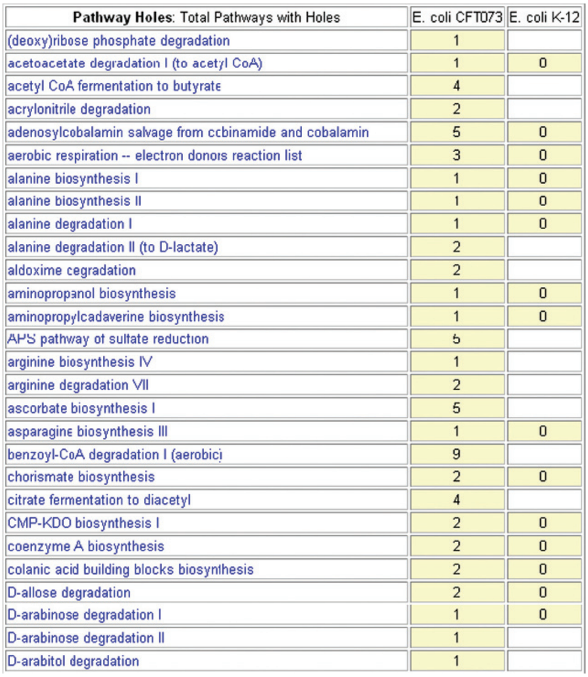
Fig. 6.5 This table presents statistics on the number of pathways present in each pathway class for the two *E. coli* strains under comparison. The two largest top-level classes, Biosynthesis and Degradation/Utilization/Assimilation, are broken down further to show the distribution of pathways among their next-level subclasses. The vast majority of pathways are assigned to only a single class. However, a small number may be assigned to more than one class



Pathway Holes	E. col CFT073	E. coli K-12
Number of Pathway Holes	328	22
Pathway Holes as a percentage of total reactions in pathways	43%	3%
Pathways with No Holes	68	190
Pathways with 1 Hole	64	9
Pathways with 2 Holes	31	1
Pathways with 3 Holes	23	2
Pathways with 4 Holes	13	0
Pathways with 5 Holes	14	0
Pathways with > 5 Holes	10	1
Total Pathways with Holes	'55	13

Fig. 6.6 A pathway hole is a reaction in a pathway for which no corresponding enzyme has been identified in the genome. Pathway holes may exist for a number of reasons: They may represent true enzymatic functions in the organism for which the gene has not yet been found, or they could represent false positive pathway predictions or cases in which the pathway in this organism differs slightly from the reference pathway in MetaCyc. This table counts all the pathway holes in each organism, and classifies pathways based on their number of pathway holes

a new table showing a list of all data elements within the columns of that row. For example, clicking on the row heading "Total Pathways with Holes" in Fig. 6.6 will produce Fig. 6.7, which shows every pathway containing at least one pathway hole (a reaction that has no assigned enzyme) in these *E. coli* strains.



Pathway Holes: Total Pathways with Holes	E. coli CFT073	E. coli K-12
(deoxy)ribose phosphate degradation	1	
acetoacetate degradation I (to acetyl CoA)	1	0
acetyl CoA fermentation to butyrate	4	
acrylonitrile degradation	2	
adenosylcobalamin salvage from ccbinamide and cobalamin	5	0
aerobic respiration -- electron donors reaction list	3	0
alanine biosynthesis I	1	0
alanine biosynthesis II	1	0
alanine degradation I	1	0
alanine degradation II (to D-lactate)	2	
aldoxime cegradation	2	
aminopropanol biosynthesis	1	0
aminopropylcadaverine biosynthesis	1	0
APS pathway of sulfate reduction	5	
arginine biosynthesis IV	1	
arginine degradation VII	2	
ascorbate biosynthesis I	5	
asparagine biosynthesis III	1	0
benzoyl-CoA degradation I (aerobic)	9	
chorismate biosynthesis	2	0
citrate fermentation to diacetyl	4	
CMP-KDO biosynthesis I	2	0
coenzyme A biosynthesis	2	0
colanic acid building blocks biosynthesis	2	0
D-allose degradation	2	0
D-arabinose degradation I	1	0
D-arabinose degradation II	1	
D-arabitol degradation	1	

Fig. 6.7 A listing of pathways in two *E. coli* PGDBs that contain pathway holes. The listing is truncated

6.3 Significance for Computational Biology

By computational biology we mean analysis of biological systems using computational methods; 51% of our survey responders said they use EcoCyc for computational biology, such as in the following areas.

6.3.1 Significance for Microbial Genome Analysis

A flood of nucleotide sequence data from microbial genomes is upon us. The genomes of more than 500 microorganisms—cultured and uncultured—have been completely sequenced, and many more will be completed in the next 5 years. Accurate, extensive analysis of these data is essential to permit them to be fully exploited in applications in medicine and biotechnology.

EcoCyc allows microbial-genome projects to produce more accurate annotations of sequenced genes, and to predict the metabolic pathways of their organisms. When gene function predictions are performed using sequence-similarity programs such as BLAST and FASTA, newly sequenced microbial genes often show similarity to *E. coli* genes. Researchers turn to EcoCyc as a source of information about *E. coli* gene function because EcoCyc is updated so frequently with literature-based information. Because *E. coli* is the genome with the highest fraction of its gene functions established experimentally, annotators for other microbial genomes are well advised to prefer sequence-similarity matches to *E. coli* genes over matches with similar scores from other organisms, to minimize the transitive annotation problem. Transitive annotation can decrease the accuracy of sequence annotation by transferring gene functions from one gene to another through long chains of similarity matches, each of which increases the likelihood of an incorrect functional prediction. Although EcoCyc curation in the 1990s focused on those genes whose products encode enzymes in metabolic pathways, it now contains rich annotations of all characterized *E. coli* genes.

In addition to predicting gene function, many scientists are using EcoCyc pathway data to predict the metabolic pathways of genomes they sequence. That prediction occurs by combining the PathoLogic module of Pathway Tools in combination with the larger MetaCyc pathway database (Caspi et al. 2008). Twice per year, SRI propagates updates to EcoCyc metabolic pathways and enzymes to MetaCyc. MetaCyc version 12.0 describes 1,036 experimentally elucidated pathways from 1,108 organisms. PathoLogic predicts the pathways of an organism by matching enzymes in the organism's annotated genome against enzymes in MetaCyc pathways, to predict which pathways from MetaCyc are present in the organism. To date, 1,300 groups have licensed Pathway Tools and MetaCyc from SRI, and tell us they are applying the software to at least 200 genomes.

As antibiotic-resistant bacteria become more prevalent, pharmaceutical companies are seeking novel microbial drug targets. Some companies are targeting enzymes within metabolic pathways (Karp 1997, 2003). Because EcoCyc improves our ability to predict the metabolic pathways of a microbe from its genomic sequence, it facilitates development of new pharmaceuticals (Karp 1997, 2003, Karp et al. 1999), such as its use by Bristol-Myers Squibb to find drug targets in *Streptococcus pneumoniae* (Thanassi et al. 2002).

6.3.2 Significance for Global Biological Studies

Because the EcoCyc data are structured within a sophisticated ontology that is amenable to computational analyses, EcoCyc allows scientists to ask questions spanning the entire genome of *E. coli*, the known metabolic network of *E. coli*, the known transport complement of *E. coli*, the known genetic regulatory network of *E. coli*, and combinations thereof. A surprisingly diverse array of systems biology studies is being fueled by EcoCyc: 40% of our survey responders said they use EcoCyc for large systematic biological studies. As we add new types of data to EcoCyc, we facilitate new types of global studies. For example, addition of new types of regulatory mechanisms will accelerate global studies of these mechanisms.

EcoCyc was used to develop methods for computing shortest path lengths within metabolic networks. These methods were used to study the topological organization of the *E. coli* metabolic network (Ravasz et al. 2002), and to investigate correlations between path lengths and factors such as genome distance between enzymes (Arita 2004, Simeonidis et al. 2003).

EcoCyc was used in several studies relating protein structure to the metabolic network. One study compared the small-molecule metabolism enzymes of yeast and *E. coli* to see which were conserved (Jardine et al. 2002). Two related studies surveyed the structural anatomy of EcoCyc pathways (Teichmann et al. 2001a,b). Two studies considered the organization of *E. coli* metabolic enzymes into protein families using EcoCyc (Rison and Thornton 2002, Tsoka and Ouzounis 2001). EcoCyc was used as a source of information on metabolic enzymes in a study that correlated sequence and functional relatedness in enzymes (Pellegrini et al. 1999).

EcoCyc was used as a source of transcriptional regulatory network information for analysis of genome-wide transcriptional regulatory networks (Ma et al. 2004), and was used to understand patterns in transcriptional control (Shen-Orr et al. 2002). EcoCyc pathways were used as a source of functionally related proteins for a study of the correlation between protein levels—evaluated based on codon bias—and functional relationship (Lithwick and Margalit 2005).

Van Dien et al. drew on EcoCyc to interpret label-tracing experiments in *Methylobacterium extorquens* to estimate flux rates through its metabolic network (Van Dien et al. 2003). Cases et al. used EcoCyc to investigate the fraction of the genome devoted to transcription-related proteins, small-molecule metabolism enzymes, and transport, for 60 bacterial genomes classified by lifestyle (Cases et al. 2003). Peregrin-Alvarez et al. used EcoCyc to study the phylogenetic extent of metabolic enzymes and pathways throughout all taxonomic domains (Peregrin-Alvarez et al. 2003).

6.4 Significance for Bioinformatics Research

The development of many new bioinformatics methods requires high-quality gold-standard datasets for training and validation of those methods; 21% of our survey responders said they use EcoCyc as a gold-standard dataset for developing bioinformatics algorithms, and 58% said they use EcoCyc for bioinformatics. As we add new types of data to EcoCyc, we facilitate development of new bioinformatics methods, for example, addition of new types of regulatory mechanisms will enable development of new predictors for those types of regulation.

Genome context methods for predicting gene function, such as phylogenetic profiles, conserved chromosomal adjacency, and the Rosetta Stone method, have been one of the major developments in bioinformatics in the last 5 years. EcoCyc played a key role in their development (Bowers et al. 2004, Enault et al. 2003, von Mering et al. 2003). EcoCyc was used to determine whether proteins that appear to share regulatory sequences might be functionally related (Studholme et al. 2004).

EcoCyc data were used to develop computational methods for predicting other key biological relationships, such as protein-protein interactions (Bowers et al. 2004, Tsoka and Ouzounis 2000), and to compute correspondences among atoms in reactants and products in biochemical reactions (Arita 2003).

EcoCyc was used as a gold standard for developing analytic and predictive computer programs. It has been used in operon prediction (Price et al. 2005, Romero and Karp 2004, Steinhauser et al. 2004) as well as for predicting promoters and transcription start sites (Burden et al. 2005, Gordon et al. 2003). EcoCyc was used as the source of metabolic pathways for genome-wide prediction of protein functions and interactions (Marcotte et al. 1999). The EcoCyc class hierarchy was used to categorize proteins for generating phylogenetic profiles (Pellegrini et al. 1999).

EcoCyc was consulted for compound-related information in a C-14-glucose radio-labeling study that followed the time dependence of various metabolite pools (Tweeddale et al. 1999). EcoCyc proved useful for investigating details of various proteins in a project to construct a whole-cell simulation (Tomita et al. 1999).

6.5 Significance for Model-Organism Database Development

In addition to *E. coli* serving as a model organism for microbial research, EcoCyc has become a model for development of bioinformatics database development for other organisms. The Pathway Tools software underlying EcoCyc is now being used in the development of many other organism-specific databases. Web links to these databases can be found at http://BioCyc.org.

Databases include

- Microbes: *Saccharomyces cerevisiae, Candida albicans, Streptomyces coelicolor, Pseudomonas aeruginosa, Rhizobium etli, Brucella suis, Coxiella burnetii, Rickettsia typhi*
- Plants: *Arabidopsis thaliana, Medicago truncatula,* multiple *Solanaceae species*
- Mammals: *Mus musculus*

6.6 Significance for Metabolic Engineering

Metabolic engineers alter microbes to produce biofuels, to produce flavor enhancers in food, to increase efficiency of production of bioproducts such as amino acids and vitamins, to produce pharmaceuticals, and to degrade toxic pollutants (Bailey 1991, Stephanopoulos and Vallino 1991). The Department of Energy GTL Project seeks to engineer microbes to solve problems of global carbon sequestration and environmental remediation (Frazier et al. 2003). The late Jay Bailey described many metabolic-engineering case studies in which heterologous proteins are introduced into cells to alter their metabolism (Bailey 1991). He wrote "No universal principles have emerged from metabolic engineering research to guide the choice of the next useful genetic alteration… there is no substitute for knowledge of the pathways involved, their regulation, and their kinetics" (Bailey 1991). Metabolic engineers consult EcoCyc and MetaCyc to select the optimal enzyme for an engineering problem, to predict undesirable side effects of a metabolic alteration, and to predict the metabolic network of their workhorse organism using Pathway Tools; 25% of our survey responders said they use EcoCyc for metabolic engineering.

The Palsson group has drawn heavily from EcoCyc to prepare quantitative flux balance models of the *E. coli* metabolic network (Edwards and Palsson 2000, Reed and Palsson 2003, Reed et al. 2003). We have recently collaborated with the Palsson group to further develop new versions of our respective models of the network (Feist et al. 2007). The Palsson group also used EcoCyc to validate results from *in silico* modeling of genome-scale *E. coli* metabolism (Reed and Palsson 2004). Other metabolic engineering studies making use of EcoCyc include (Chassagnole et al. 2002, Jardine et al. 2002, Weber et al. 2002)

6.7 Significance for Education

Of our survey responders, 20% said they use EcoCyc in graduate or undergraduate classes that they teach. The classes include Metabolic Network Analysis; Microbial Physiology; Introduction to Bioinformatics; Molecular Genetics; Genomics, Proteomics and Systems Biology; and Microbial Biotechnology. Dr. R. Gunsalus of the University of California Los Angeles is developing a Web portal to EcoCyc for use in undergraduate microbiology education.

Acknowledgments EcoCyc has greatly benefited from the involvement and guidance of M. Riley, J. Ingraham, R. Gunsalus, I. Paulsen, and J. Collado-Vides. The graph-tracks capability was suggested by R. Landick and implemented by J. Myers. EcoCyc is supported by grant GM077678 from the National Institutes of Health. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

References

Arita M (2003) *In silico* atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. Genome Res 13(11):2455–66

Arita M (2004) The metabolic world of *Escherichia coli* is not small. Proc Natl Acad Sci USA 101(6):1543–7

Bailey JE (1991) Toward a science of metabolic engineering. Science 252(5013):1668–75

Bowers PM, Pellegrini M, Thompson MJ et al. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5(5):R35

Burden S, Lin YX, Zhang R (2005) Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. Bioinformatics 21(5):601–7

Cases I, de Lorenzo V, Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. Trends Microbiol 11(6):248–53

Caspi R, Foerster H, Fulcher CA et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36(Database issue):D623–31

Chassagnole C, Letisse F, Diano A et al. (2002) Carbon flux analysis in a pantothenate overproducing *Corynebacterium glutamicum* strain. Mol Biol Rep 29(1–2):129–34

Edwards JS, Palsson BO (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA 97(10):5528–33

Enault F, Suhre K, Poirot O et al. (2003) Phydbac (phylogenomic display of bacterial genes): An interactive resource for the annotation of bacterial genomes. Nucleic Acids Res 31(13): 3720–2

Feist AM, Henry CS, Reed JL et al. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121

Frazier ME, Johnson GM, Thomassen DG et al. (2003) Realizing the potential of the genome revolution: the genomes to life program. Science 300(5617):290–3

Gama-Castro S, Jimenez-Jacinto V, Peralta-Gil M et al. (2008) RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. Nucleic Acids Res 36(Database issue):D120–4

Gordon L, Chervonenkis AY, Gammerman AJ et al. (2003) Sequence alignment kernel for recognition of promoter regions. Bioinformatics 19(15):1964–71

Jardine O, Gough J, Chothia C et al. (2002) Comparison of the small molecule metabolic enzymes of *Escherichia coli* and *Saccharomyces cerevisiae*. Genome Res 12(6):916–29

Karp PD (1997) Use of metabolic databases to guide target selection for anti-microbial drug design. Blackwell Science Ltd., Oxford, UK

Karp PD (2003) The Pathway Tools software and its role in anti-microbial drug discovery. Marcel Dekker, Inc., New York

Karp PD, Keseler IM, Shearer A et al. (2007) Multidimensional annotation of the *Escherichia coli* K-12 genome. Nucleic Acids Res 35(22):7577–90

Karp PD, Krummenacker M, Paley S et al. (1999) Integrated pathway-genome databases and their role in drug discovery. Trends Biotechnol 17(7):275–81

Karp PD, Ouzounis CA, Moore-Kochlacs C et al. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 33(19):6083–9

Lithwick G, Margalit H (2005) Relative predicted protein levels of functionally associated proteins are conserved across organisms. Nucleic Acids Res 33(3):1051–7

Ma HW, Kumar B, Ditges U et al. (2004) An extended transcriptional regulatory network of *Escherichia coli* and analysis of its hierarchical structure and network motifs. Nucleic Acids Res 32(22):6643–9

Marcotte EM, Pellegrini M, Thompson MJ et al. (1999) A combined algorithm for genome-wide prediction of protein function. Nature 402(6757):83–6

Pellegrini M, Marcotte EM, Thompson MJ et al. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci USA 96(8): 4285–8

Peregrin-Alvarez JM, Tsoka S, Ouzounis CA (2003) The phylogenetic extent of metabolic enzymes and pathways. Genome Res 13(3):422–7

Price MN, Huang KH, Alm EJ et al. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res 33(3):880–92

Ravasz E, Somera AL, Mongru DA et al. (2002) Hierarchical organization of modularity in metabolic networks. Science 297(5586):1551–5

Reed JL, Palsson BO (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. J Bacteriol 185(9):2692–9

Reed JL, Palsson BO (2004) Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome Res 14(9):1797–805

Reed JL, Vo TD, Schilling CH et al. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). Genome Biol 4(9):R54

Rison SC, Thornton JM (2002) Pathway evolution, structurally speaking. Curr Opin Struct Biol 12(3):374–82

Romero PR, Karp PD (2004) Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. Bioinformatics 20(5):709–17

Salgado H, Gama-Castro S, Martinez-Antonio A et al. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. Nucleic Acids Res 32(Database issue):D303–6

Shen-Orr SS, Milo R, Mangan S et al. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31(1):64–8

Simeonidis E, Rison SC, Thornton JM et al. (2003) Analysis of metabolic networks using a pathway distance metric through linear programming. Metab Eng 5(3):211–9

Steinhauser D, Junker BH, Luedemann A et al. (2004) Hypothesis-driven approach to predict transcriptional units from gene expression data. Bioinformatics 20(12):1928–39

Stephanopoulos G, Vallino JJ (1991) Network rigidity and metabolic engineering in metabolite overproduction. Science 252(5013):1675–81

Studholme DJ, Bentley SD, Kormanec J (2004) Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor*. BMC Microbiol 4(14):14

Teichmann SA, Rison SC, Thornton JM et al. (2001a) The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. J Mol Biol 311(4):693–708

Teichmann SA, Rison SC, Thornton JM et al. (2001b) Small-molecule metabolism: an enzyme mosaic. Trends Biotechnol 19(12):482–6

Thanassi JA, Hartman-Neumann SL, Dougherty TJ et al. (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. Nucleic Acids Res 30(14):3152–62

Tomita M, Hashimoto K, Takahashi K et al. (1999) E-CELL: software environment for whole-cell simulation. Bioinformatics 15(1):72–84

Tsoka S, Ouzounis CA (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. Nat Genet 26(2):141–2

Tsoka S, Ouzounis CA (2001) Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*. Genome Res 11(9):1503–10

Tweeddale H, Notley-McRobb L, Ferenci T (1999) Assessing the effect of reactive oxygen species on *Escherichia coli* using a metabolome approach. Redox Rep 4(5):237–41

Van Dien SJ, Strovas T, Lidstrom ME (2003) Quantification of central metabolic fluxes in the facultative methylotroph *Methylobacterium extorquens* AM1 using 13C-label tracing and mass spectrometry. Biotechnol Bioeng 84(1):45–55

von Mering C, Zdobnov EM, Tsoka S et al. (2003) Genome evolution reveals biochemical networks and functional modules. Proc Natl Acad Sci USA 100(26):15428–33

Weber J, Hoffmann F, Rinas U (2002) Metabolic adaptation of *Escherichia coli* during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. Biotechnol Bioeng 80(3):320–30