D464–D470 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 30 October 2008*
doi:10.1093/nar/gkn751

EcoCyc: A comprehensive view of *Escherichia coli* biology

Ingrid M. Keseler[1], César Bonavides-Martínez[2], Julio Collado-Vides[2], Socorro Gama-Castro[2], Robert P. Gunsalus[3], D. Aaron Johnson[4], Markus Krummenacker[1], Laura M. Nolan[5], Suzanne Paley[1], Ian T. Paulsen[5], Martin Peralta-Gil[2], Alberto Santos-Zavaleta[2], Alexander Glennon Shearer[1] and Peter D. Karp[1,*]

[1]SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025, [2]Program of Computational Genomics, Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, AP 565-A, Cuernavaca, Morelos 62100, Mexico, [3]Department of Microbiology, Immunology, and Molecular Genetics and the Molecular Biology Institute, University of California, Los Angeles, CA 90095, [4]J. Craig Venter Institute, Rockville, MD 20850, USA and [5]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia 2109

Received September 15, 2008; Accepted October 5, 2008

ABSTRACT

EcoCyc (http://EcoCyc.org) provides a comprehensive encyclopedia of *Escherichia coli* biology. EcoCyc integrates information about the genome, genes and gene products; the metabolic network; and the regulatory network of *E. coli*. Recent EcoCyc developments include a new initiative to represent and curate all types of *E. coli* regulatory processes such as attenuation and regulation by small RNAs. EcoCyc has started to curate Gene Ontology (GO) terms for *E. coli* and has made a dataset of *E. coli* GO terms available through the GO Web site. The curation and visualization of electron transfer processes has been significantly improved. Other software and Web site enhancements include the addition of tracks to the EcoCyc genome browser, in particular a type of track designed for the display of ChIP-chip datasets, and the development of a comparative genome browser. A new Genome Omics Viewer enables users to paint omics datasets onto the full *E. coli* genome for analysis. A new advanced query page guides users in interactively constructing complex database queries against EcoCyc. A Macintosh version of EcoCyc is now available. A series of Webinars is available to instruct users in the use of EcoCyc.

OVERVIEW OF EcoCyc CONTENT

Since the last NAR Database Issue publication on EcoCyc four years ago (1), significant additions and changes to the content and features of EcoCyc have occurred. EcoCyc staff perform an ongoing literature-based curation of the *Escherichia coli* genome, whose methodology and results were described in detail in 2007 (2). The EcoCyc curators edit gene names and functions, and write mini-reviews about each *E. coli* gene product and multimeric complex. These mini-reviews include extensive citations to the experimental literature. In mid-2006, EcoCyc reached an important milestone when EcoCyc curators had performed literature searches for every *E. coli* gene and had written mini-reviews for every gene for which experimental literature was found. In EcoCyc 12.5, released during fall 2008, 2650 (59.3%) *E. coli* genes have experimentally defined functions. Table 1 provides an overview of the current contents of EcoCyc.

RECENT INITIATIVES

Electron transfer enzymes and associated pathways in the membrane

Previously, curation of electron transfer reactions in EcoCyc was limited to brief written summaries of the gene products and protein complexes. This approach did not provide for a visual representation of the electron transfer enzymes in the membrane, nor did it indicate

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

Table. 1. Overview of the current contents of EcoCyc

Data type	Number
Genes	4472
Gene products covered by a mini-review	3532
Gene products with experimental evidence for their functions	2650
Enzymes	1397
Metabolic reactions	1326
Compounds	1362
Transporters	241
Transport reactions	267
Transported substrates	182
Transcription factors	196
Transcription units	3359
With experimental evidence	940
Regulatory interactions	4792
Transcription initiation	2417
Transcription attenuation	18
Enzyme modulation	2342
Other (includes regulation by sRNAs)	15
Literature citations	17258

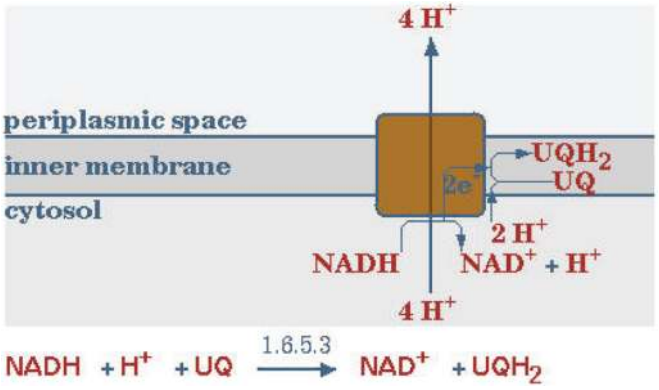

Figure 1. Graphical representation of the electron transfer reaction catalyzed by NADH:ubiquinone oxidoreductase I (NDH-1). The cellular location of the reaction substrates, electron flow from the cytoplasmic substrates to the membrane-localized ubiquinone and vectorial proton transport across the membrane are shown. A summary reaction equation corresponding to the Enzyme Commission nomenclature (which does not include proton and electron flow) is shown below the graphical representation.

known or potential roles in cellular electron transfer and proton movement relative to the cell compartments. To address these issues, we have extended the Pathway Tools software that underlies EcoCyc in two respects: First, it can now visually depict electron transfer enzyme complexes and their associated balanced oxidation/reduction reactions (Figure 1). Reaction displays now show enzyme membrane localization, the flow of all substrates and products, and the fate of the protons associated with the overall reactions. Second, the software can now depict electron transfer pathways that consist of coupled systems of electron transfer enzymes (Figure 2).

E. coli possesses more than 25 enzymes and enzyme complexes that participate in the oxidation of primary electron donors or in the reduction of terminal electron acceptors during different cell culture conditions. The literature-based curation for approximately 15 electron transfer enzymes and enzyme complexes has been updated, and associated membrane depictions and balanced reactions are available. Electron transfer pathways have been generated and curated for 10 sets of electron donor/acceptor pairs.

An example of a membrane depiction is shown in Figure 1 for the *E. coli* enzyme NADH dehydrogenase I, encoded by the *nuoABCDEFGHIJKLMN* operon. Herein, the oxidation of NADH is shown to occur at the cytoplasmic face of the enzyme with electron transfer within the enzyme to the physiological electron acceptors, ubiquinone (UQ) or menaquinone (MQ).

Combining the oxidation reactions for a physiological electron donor and an acceptor yields an electron transport pathway. For example, in Figure 2 the NADH dehydrogenase I enzyme shown in Figure 1 is combined with cytochrome *bo* oxidase (*cyoABCD*) to represent the transfer of electrons from NADH to molecular oxygen (O_2). Net movement of protons across the membrane by each enzyme complex provides, in part, the proton motive force (PMF) needed for ATP synthesis.

Updates to regulation of transcription initiation

Curation of transcriptional regulation is performed by the RegulonDB group at the Center for Genomic Sciences, Universidad Nacional Autónoma de México. Curation of older literature on transcriptional regulation was completed in December 2006 and since then, data from new literature is consistently added to EcoCyc shortly after publication.

After reports of differences and apparent inconsistencies between the transcriptional regulatory networks of EcoCyc and RegulonDB appeared (3,4), we undertook detailed curation that led to fully synchronized content and releases in both databases (5). Other systematic curation efforts included the sigmulons of σ^{54} (RpoN), σ^{28} (FliA), σ^{19} (FecI), σ^{24} (RpoE), σ^{32} (RpoH), and σ^{38} (RpoS); various metabolic and motility regulons; and representations of the binding sites for the ArcA and NarL transcription factors. In addition, we have developed guidelines for transcription factor summaries to include relevant physiological data found in the literature that cannot be easily added as database objects. Many summaries have been updated according to these guidelines.

To facilitate the tracking and querying of data based on the quality of the evidence, we have classified the types of evidence used to annotate regulatory objects as 'strong' or 'weak'. Strong evidence corresponds to experiments—irrespective of methodology—that provide direct physical evidence. Examples of strong evidence include the experimental mapping of transcription start sites and DNA binding of purified transcription factors. Evidence such as that from gene expression analyses that provide only indirect evidence is considered weak. Strong and weak evidence types are graphically distinguished by using solid or dashed lines for the corresponding objects (such as promoter arrows).

To expand the information about transcription regulation of *E. coli*, the RegulonDB group has incorporated

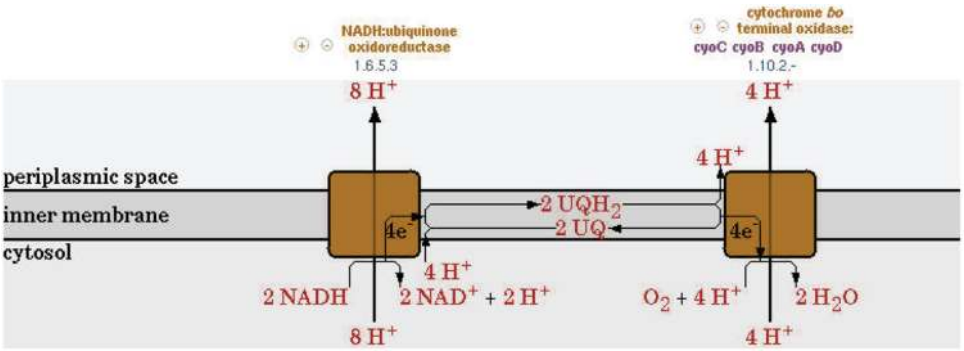

Figure 2. Graphical representation of the electron transfer pathway between the NDH-1 and cytochrome *bo* terminal oxidase enzyme complexes. The pathway couples the oxidation of NADH to the reduction of oxygen to water via transfer of electrons in the membrane by ubiquinone/ubiquinol and includes proton transport across the membrane by both enzymes, which generates a proton-motive force.

Figure 3. Graphical representation of regulation of the *thrLABC* operon by transcriptional attenuation. When the charged tRNAs L-isoleucyl-$tRNA^{Ile}$ or L-threonyl-$tRNA^{Thr}$ are available in the cell, the *thrL* open reading frame is translated freely, leading to the formation of the attenuator secondary structure (represented by a lollipop icon) and termination of transcription by RNA polymerase. The negative effect on transcription is indicated by red color of the icons for the charged tRNAs, and the termination of transcription is indicated by X at the location of the attenuator.

various new types of experimental and predicted data into EcoCyc. A collection of 259 new transcription start sites, which resulted from a high-throughput experimental modified RACE approach, was added (6). Promoters and DNA binding sites with evidence from at least two types of high-throughput data (such as computational predictions, microarrays and ChIP-chip experiments) have been added to EcoCyc. Examples include a collection of 54 σ^{32} promoters experimentally identified by ChIP-chip and by gene expression assays (7); 45 σ^{32} promoters identified by microarray analysis, transcription initiation mapping and computational analysis (8); and 45 Fur DNA binding sites identified by computational prediction and binding of purified protein (9).

Beyond regulation of transcription initiation

EcoCyc has included information about the regulation of both transcription initiation and enzyme activity for many years. A major new EcoCyc initiative is to expand the database schema and content to include other types of regulation, such as attenuation and regulation of translation by small RNAs (sRNAs). For example, the EcoCyc schema can now represent all six known types of regulation by attenuation of transcription, each of which involves slightly different database fields to capture aspects such as the regulatory ligand, protein and RNA regions involved. This initiative will provide both more complete information about *E. coli* regulation and the regulatory datasets that can be used by bioinformaticians to develop predictors for a broader diversity of regulatory interactions from genome datasets.

All known examples of ribosome-mediated attenuation in the pathways of amino acid biosynthesis have been added to EcoCyc in release 12.5. For example, Figure 3 shows regulation of the *thrLABC* operon by attenuation, which is modulated by the availability of charged isoleucyl- and threonyl-tRNA. In this example of attenuation, translation of the *thrL* leader peptide open reading frame influences the formation of an attenuator structure. When charged isoleucyl- and threonyl-tRNAs are abundant, unobstructed translation by the ribosome enables the formation of a secondary structure that acts as a terminator, releasing RNA polymerase and halting transcription of the operon. On the EcoCyc display, the charged tRNAs are represented as rods. Their role in modulating termination at the attenuator is indicated by their red color and the 'X' near the terminator structure; this shows at a glance that a charged tRNA leads to premature termination. Curation of other attenuation systems is ongoing.

An example of the representation of regulation by sRNAs is shown in Figure 4. The transcription unit that encompasses the *glmUS* operon is shown. Expression of this operon is regulated at the level of transcription initiation by the transcription factor NagC (10), whose binding sites are shown as green boxes upstream of the *glmUS* transcription start site. In addition, the sRNA GlmZ was recently shown to regulate translation of the second open reading frame, *glmS* (11,12). *glmS* encodes L-glutamine:D-fructose-6-phosphate aminotransferase, the

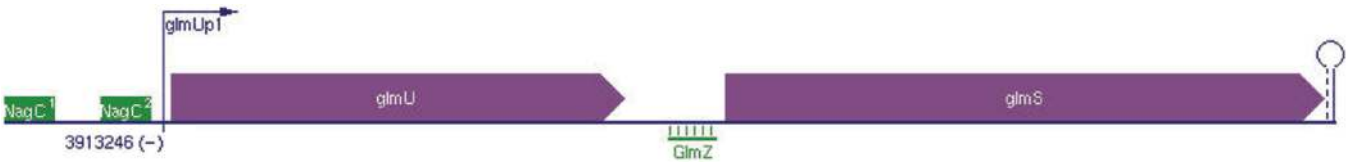

Figure 4. Graphical representation of the regulation of *glmUS* expression by the transcription factor NagC and the small RNA GlmZ. Transcription initiation at the *glmUS* promoter is positively regulated by NagC, and is represented by green boxes at the NagC binding sites. The small RNA GlmZ positively regulates translation of the *glmS* open reading frame via base-pairing, which is indicated by short lines connecting GlmZ with the mRNA.

enzyme that catalyzes the first step in the biosynthesis of UDP-*N*-acetylglucosamine, which is used as the precursor for the synthesis of peptidoglycan, lipid A and the enterobacterial common antigen. Genetic experiments suggest that full-length GlmZ interacts directly with the 5′ UTR of *glmS*, unmasking the ribosome binding site and thus activating translation (11,12). The interaction of GlmZ with the *glmUS* mRNA is shown by a bar (representing GlmZ) that is connected with lines to *glmUS*, suggesting base-pairing at the position indicated.

The 12.5 release of EcoCyc contains 19 examples of attenuation and 15 examples of regulation by mechanisms other than transcription initiation, attenuation, or regulation of enzyme activity. We are actively expanding both the curation of the preceding regulatory mechanisms and the ability of the Pathway Tools software to handle additional regulatory mechanisms.

Annotation of EcoCyc gene products with Gene Ontology terms

Gene Ontology (GO) is an accepted standard for ontological annotation of gene products (www.Gene Ontology.org). The EcoCyc project has been annotating *E. coli* genes with GO terms for the past two years. Overall, the more than 38 000 GO terms present in EcoCyc have been derived from four sources: (i) GO terms were inferred from a mapping from the original MultiFun (13) ontology annotations within EcoCyc to GO terms; (ii) GO terms were inferred from a mapping from the Enzyme Commission (EC) numbers present within EcoCyc to GO terms; (iii) GO term assignments are manually curated by EcoCyc curators on an ongoing basis; and (iv) many GO terms were imported into EcoCyc from UniProt. EcoCyc and the EcoliWiki project (www.EcoliWiki.net) are jointly producing an official data file of *E. coli* GO terms that we regularly submit to the GO project, and that is available from the GO Web site at http://www.geneontology.org/GO.current.annotations.shtml.

GO terms are found on EcoCyc gene and gene product pages and provide a useful way of finding all *E. coli* genes with a common function. For example, *rsmD* encodes an rRNA methyltransferase and is annotated with the GO process term for rRNA methylation, GO:0031167. Clicking that GO term navigates the user to a page that both provides the definition of that GO term and lists all other gene products within EcoCyc that have been annotated with that GO term. The GO term annotations within EcoCyc should be considered incomplete, as manual curation of GO terms is ongoing.

Updates to metabolic pathways

Although EcoCyc has now expanded far beyond its initial role, EcoCyc began as a database of *E. coli* metabolism, primarily describing metabolic enzymes and pathways. Therefore, annotations for many metabolic enzymes are among the oldest entries in EcoCyc. During the past decade, significant progress has been made in understanding *E. coli* metabolic pathways and their enzymes. Therefore, we have begun to systematically re-annotate these pathways; in release 12.5, 41 pathways that were entered into EcoCyc more than ten years ago, as well as 19 more recently added pathways, have been updated. As part of this effort, the curation of more than 180 metabolic enzymes has already been updated to reflect the latest state of knowledge.

NEW SOFTWARE AND WEB SITE FEATURES

Genome browser tracks

The EcoCyc genome browser now supports a track mechanism to aid users in visually analyzing positional datasets with respect to genome features such as the positions of genes, promoters and transcription factor binding sites. Examples include datasets of predicted promoters, predicted transcription factor binding sites and ChIP-chip datasets. Datasets encoded as GFF-format files (http://www.sanger.ac.uk/Software/formats/GFF/) can be loaded into the desktop or Web versions of EcoCyc. Figure 5 shows a type of track specifically designed for the visualization of ChIP-chip data called a graph track.

Multi-genome browser

Users of EcoCyc include both researchers who study the biology of *E. coli* and those who use *E. coli*, and thus EcoCyc, as a reference for their research in other organisms. To support both types of users, we have added several comparative tools to EcoCyc. The comparative genome browser is accessible from every gene page, and allows a user to select organisms from the hundreds that are available via the BioCyc database collection at BioCyc.org (14) and to then examine the ortholog of the starting gene in its local context within each selected organism. For example, Figure 6 shows the *E. coli* gene *thrA* aligned with its orthologs in several other organisms. The starting gene is marked with hash marks and aligned

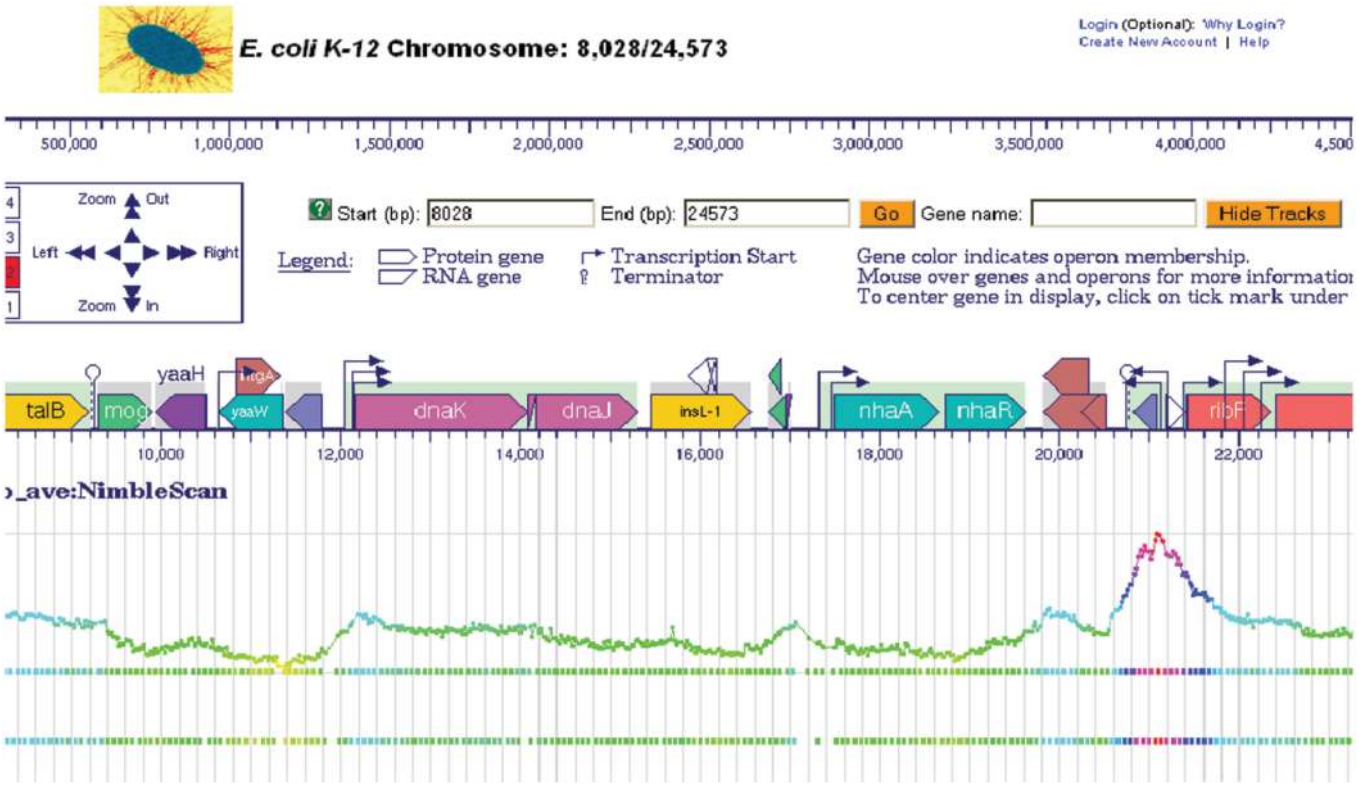

Figure 5. The track capabilities of the genome browser. Three tracks are shown below the depicted genes and promoters. The first (highest) of the three tracks is a graph track that is designed to depict ChIP-chip datasets. Graph tracks plot positional information against the genome, and include a Y-component that depicts an intensity value associated with a genome region. Each point in the graph track depicts a ChIP-chip measurement at a given genome region. Its intensity is shown both in the height dimension and in the selected color. Below the graph track are two horizontal tracks that depict the same ChIP-chip dataset, but without the height information–intensities are depicted with color only. Two horizontal tracks are shown to eliminate visual overlap between the regions.

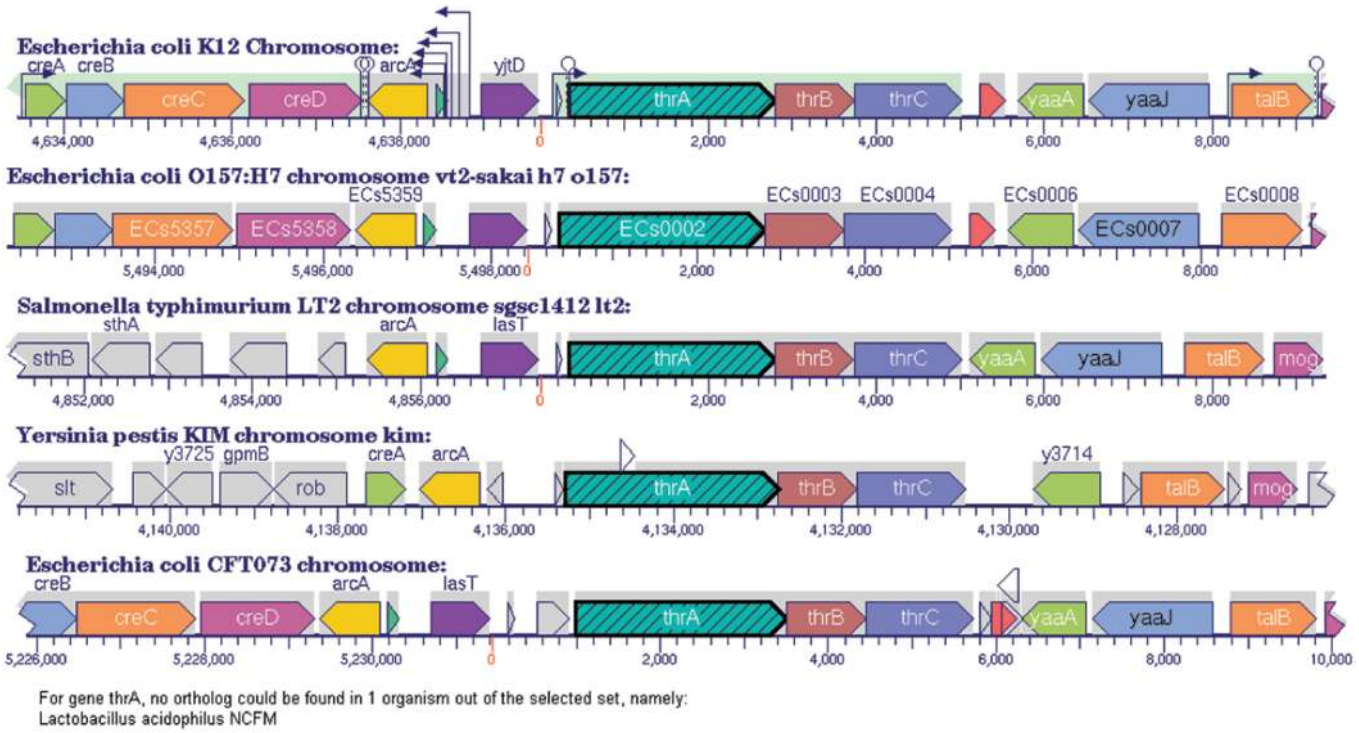

Figure 6. The multi-genome browser displays a gene and its orthologs across multiple organisms. Hash marks show the starting gene of interest. Orthologs—including, but not limited to, the starting gene—are marked by colors.

Figure 7. The Genome Omics Viewer enables the analysis of large-scale data sets in the context of the entire *E. coli* genome. Colors show gene expression based on user-defined benchmarks. Membership in transcription units is indicated by the lines below the genes. Users can click through from individual genes to the appropriate gene page within EcoCyc.

across the displays. Note that the other orthologs present are marked with the same color. For example, the adjacent gene *thrB* has an ortholog present in each organism displayed. The tool also indicates at the bottom of the page when no ortholog could be found. Using the multi-genome browser, users can query a broad range of organisms in search of orthologs and then can see the extent to which those orthologs have maintained their genetic context relative to *E. coli*.

The Genome Omics viewer

Many users come to EcoCyc with large-scale datasets that include gene expression, proteomic and metabolomic data. As described in our earlier report on the EcoCyc database, these datasets can be viewed in the context of the *E. coli* metabolic network via the Cellular Omics Viewer, which is a tool that enables users to 'paint' the results from these datasets onto the Cellular Overview diagram. To this tool, we have recently added the Genome Omics Viewer. This new viewing tool enables the display of large-scale gene-related datasets on the full *E. coli* genome, providing a valuable additional tool for the interpretation of high-throughput data. As shown in Figure 7, the Genome Omics Viewer differs from the EcoCyc Genome Browser both by providing a schematic rather than a 'to-scale' view of the genome and by placing an emphasis on operon membership and adjacent genes. In combination, the Genome and Cellular Omics Viewers enable interpretation of large datasets in both the metabolic and genomic contexts.

Advanced query page

The new EcoCyc advanced query page is accessible by clicking the 'Advanced Query Form' button located at

the bottom of most EcoCyc data pages. The resulting page enables users to interactively construct complicated, multi-criteria searches against EcoCyc. Example queries include 'Find all proteins of *E. coli* K-12 for which the DNA-FOOTPRINT-SIZE is smaller than 10' and 'Find all proteins of *E. coli* K-12 containing more than one subunit and that catalyze a reaction in which pyruvate is a substrate'. Instructions for the advanced query page are available at http://www.biocyc.org/webQueryDoc.html.

Desktop EcoCyc now available for the Macintosh

For many years we have provided a version of EcoCyc that runs as an application on a user's local laptop or workstation computer. This form of EcoCyc access is highly recommended for frequent EcoCyc users because it provides faster execution and more capabilities than the Web version of EcoCyc. Scientists who use either the omics data analysis facilities or the genome browser tracks will find this version's faster speeds particularly useful. Differences between the desktop and Web versions of EcoCyc are summarized at http://www.biocyc.org/desktop-vs-web-mode.shtml.

In early 2008, we adapted the desktop EcoCyc software to run on the Macintosh, adding one more personal computer option to the existing PC/Windows and PC/Linux platforms.

New Web Accounts system

The EcoCyc Web site now allows users to create accounts through which they can customize the appearance of EcoCyc pages, store organism sets for comparative operations, configure default settings for the Omics Viewers, and register to receive periodic email updates about EcoCyc. See the 'Create New Account' link in the upper right corner of most EcoCyc Web pages.

Learn about EcoCyc through Webinars

We have produced several video tutorials that walk users through the basic and advanced use of the EcoCyc and BioCyc Web sites, and that cover the unique features of the desktop software. These videos are available in several formats directly from the BioCyc site (http://www.biocyc.org/webinar.shtml), and as podcasts via either iTunes (search for 'BioCyc'in the podcasts section of the iTunes Store) or the video-sharing site YouTube (http://www.youtube.com/user/SRIBRG).

AVAILABILITY

Flat files that contain the EcoCyc data are freely available for download at http://www.biocyc.org/download.shtml. The Pathway Tools software/database bundle is freely available to academic researchers.

ACKNOWLEDGEMENTS

We thank Dr Robert Landick for suggesting the graph-track display.

FUNDING

National Institutes of Health (grants GM077678 and GM75742 to P.D.K., GM071962 to J.C.-V.). Funding for open access charge: NIH grant GM077678.

Conflict of interest statement. SRI authors benefit from a commercial licensing program for Pathway Tools.

REFERENCES

1. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res.*, **33**, D334–337.
2. Karp,P.D., Keseler,I.M., Shearer,A., Latendresse,M., Krummenacker,M., Paley,S.M., Paulsen,I., Collado-Vides,J., Gama-Castro,S., Peralta-Gil,M. *et al.* (2007) Multidimensional annotation of the Escherichia coli K-12 genome. *Nucleic Acids Res.*, **35**, 7577–7590.
3. Ma,H.W., Kumar,B., Ditges,U., Gunzer,F., Buer,J. and Zeng,A.P. (2004) An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. *Nucleic Acids Res.*, **32**, 6643–6649.
4. Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat. Genet.*, **31**, 64–68.
5. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Peralta-Gil,M., Penaloza-Spinola,M.I., Martinez-Antonio,A., Karp,P.D. and Collado-Vides,J. (2006) The comprehensive updated regulatory network of Escherichia coli K-12. *BMC Bioinformatics*, **7**, 5.
6. Gama-Castro,S., Jimenez-Jacinto,V., Peralta-Gil,M., Santos-Zavaleta,A., Penaloza-Spinola,M.I., Contreras-Moreira,B., Segura-Salazar,J., Muniz-Rascado,L., Martinez-Flores,I., Salgado,H. *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**, D120–124.
7. Wade,J.T., Roa,D.C., Grainger,D.C., Hurd,D., Busby,S.J., Struhl,K. and Nudler,E. (2006) Extensive functional overlap between sigma factors in Escherichia coli. *Nat. Struct. Mol. Biol.*, **13**, 806–814.
8. Nonaka,G., Blankschien,M., Herman,C., Gross,C.A. and Rhodius,V.A. (2006) Regulon and promoter analysis of the *E. coli* heat-shock factor, sigma32, reveals a multifaceted cellular response to heat stress. *Genes Dev.*, **20**, 1776–1789.
9. Chen,Z., Lewis,K.A., Shultzaberger,R.K., Lyakhov,I.G., Zheng,M., Doan,B., Storz,G. and Schneider,T.D. (2007) Discovery of Fur binding site clusters in Escherichia coli by information theory models. *Nucleic Acids Res.*, **35**, 6762–6777.
10. Plumbridge,J. (1995) Co-ordinated regulation of amino sugar biosynthesis and degradation: the NagC repressor acts as both an activator and a repressor for the transcription of the glmUS operon and requires two separated NagC binding sites. *EMBO J.*, **14**, 3958–3965.
11. Urban,J.H. and Vogel,J. (2008) Two seemingly homologous non-coding RNAs act hierarchically to activate glmS mRNA translation. *PLoS Biol.*, **6**, e64.
12. Reichenbach,B., Maes,A., Kalamorz,F., Hajnsdorf,E. and Gorke,B. (2008) The small RNA GlmY acts upstream of the sRNA GlmZ in the activation of glmS expression and is subject to regulation by polyadenylation in Escherichia coli. *Nucleic Acids Res.*, **36**, 2570–2580.
13. Serres,M.H. and Riley,M. (2000) MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microb. Comp. Genomics*, **5**, 205–222.
14. Caspi,R., Foerster,H., Fulcher,C.A., Kaipa,P., Krummenacker,M., Latendresse,M., Paley,S., Rhee,S.Y., Shearer,A.G., Tissier,C. *et al.* (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **36**, D623–631.