Briefings in Bioinformatics, 2017, 1–9

doi: 10.1093/bib/bbx085
Paper

The BioCyc collection of microbial genomes and metabolic pathways

Peter D. Karp, Richard Billington, Ron Caspi, Carol A. Fulcher, Mario Latendresse, Anamika Kothari, Ingrid M. Keseler, Markus Krummenacker, Peter E. Midford, Quang Ong, Wai Kit Ong, Suzanne M. Paley and Pallavi Subhraveti

Corresponding author: Peter D. Karp, Bioinformatics Research Group, SRI International, 333 Ravenswood Ave, AE206, Menlo Park, CA 94025, USA.
E-mail: pkarp@ai.sri.com

Abstract

BioCyc.org is a microbial genome Web portal that combines thousands of genomes with additional information inferred by computer programs, imported from other databases and curated from the biomedical literature by biologist curators. BioCyc also provides an extensive range of query tools, visualization services and analysis software. Recent advances in BioCyc include an expansion in the content of BioCyc in terms of both the number of genomes and the types of information available for each genome; an expansion in the amount of curated content within BioCyc; and new developments in the BioCyc software tools including redesigned gene/protein pages and metabolite pages; new search tools; a new sequence-alignment tool; a new tool for visualizing groups of related metabolic pathways; and a facility called SmartTables, which enables biologists to perform analyses that previously would have required a programmer's assistance.

Key words: genome databases; microbial genome databases; metabolic pathway databases

Peter Karp is the Director of the SRI Bioinformatics Research Group (BRG).
Richard Billington is a Senior Software Engineer at SRI International; he obtained an MS degree from the University of Pennsylvania and a BS degree from the University of California at Santa Cruz. He has held research positions at the University of Michigan, University of Pennsylvania, Georgia Tech, and Sandia National Labs.
Ron Caspi has a PhD in Molecular Microbiology. He is currently the sole curator of the MetaCyc database.
Carol A. Fulcher, is a Scientific Database Curator at SRI International.
Mario Latendresse is a Computer Scientist at SRI International. He has worked on the flux balance analysis module, atom mappings, Gibbs free energies computation, BioVelo, Route Search and visualization tools.
Anamika Kothari obtained her bachelor's degree from K.C. College Mumbai; she is a Scientific Database Curator at SRI International.
Ingrid Keseler, is a Scientific Database Curator at SRI International.
Markus Krummenacker is a computer scientist. He has worked on the BioCyc genome browser, reaction and compound display pages, and electron transfer diagrams.
Peter E. Midford is a Chemo/Bioinformatics Scientist at SRI International. Before joining SRI, he worked on phylogenetic databases and phenotype ontologies for several collaborative projects.
Quang Ong received a BS in Industrial Management and is a Scientific Programmer at SRI International.
Wai Kit Ong is a Metabolic Modeler at BRG. He is responsible for building genome-scale metabolic network models of bacteria with a focus on those found in the human gut microbiome.
Suzanne Paley is a Senior Software Developer and has been with the BioCyc project since its inception. She is responsible for Pathway Collages, and many other BioCyc query, analysis and visualization tools.
Pallavi Subhraveti is the Release Manager at BRG. She is also responsible for building PGDBs in large scale.
Submitted: 18 May 2017; **Received (in revised form):** 22 June 2017

Introduction

BioCyc.org is a microbial genome Web portal that combines thousands of genomes with additional information inferred by computer programs, imported from other databases (DBs) and curated from the biomedical literature by biologist curators. BioCyc also provides an extensive range of query tools, visualization services and analysis software.

BioCyc has been developed over a 25 year period, beginning with the EcoCyc DB for *Escherichia coli*. Over time, the content of BioCyc has expanded in terms of the number of genomes, the types of information available for each genome and the amount of curated content. BioCyc has also grown to include some eukaryotic genomes (although its main emphasis is microbial).

The software behind BioCyc, called Pathway Tools [1,2], has also expanded in many ways during this period, such as to support regulatory networks, omics data analysis and metabolic modeling. Recent enhancements include redesigned gene/protein pages and metabolite pages, new search tools, a new sequence-alignment tool, a new tool for visualizing groups of related metabolic pathways and a facility called SmartTables, which enables biologists to perform analyses that previously would have required a programmer's assistance.

Expansion of BioCyc DB content

Each BioCyc DB describes one sequenced genome, with the exception of the MetaCyc DB, which describes experimentally studied metabolic pathways from all domains of life. Since 2011, BioCyc has expanded from 1000 genomes to 9300 genomes. The majority of those genomes were obtained from Genbank RefSeq and from the Human Microbiome Project complete genomes DB. As the majority of sequenced microbial genomes are of interest to a relatively small number of researchers, BioCyc emphasizes breadth and quality of information for more highly used genomes at the expense of number of genomes.

To facilitate access to the more commonly used BioCyc Pathway/Genome Databases (PGDBs), we have created the set of home pages listed in Table 1. When entering BioCyc through these home pages, the user's default organism will be set to the BioCyc PGDB for the primary strain for that species.

Workflow for generation of BioCyc PGDBs

To produce new BioCyc PGDBs, we process each BioCyc genome through the computational steps shown in Figure 1 both to computationally infer new information for the genome and to integrate additional information from other bioinformatics DBs. The amount of information found by the different import steps will vary for different organisms. Note that we retain the original genome annotation that was present in the downloaded genome file(s) for each organism.

Table 1. Home pages for BioCyc organisms

Home page	Genus
ecocyc.org	*Escherichia coli*
helicobacter.biocyc.org	*Helicobacter pylori*
vibrio.biocyc.org	*Vibrio cholerae*
listeria.biocyc.org	*Listeria monocytogenes*
salmonella.biocyc.org	*Salmonella enterica*
shigella.biocyc.org	*Shigella flexneri*
cdifficile.biocyc.org	*Clostridium difficile*
mycobacterium.biocyc.org	*Mycobacterium tuberculosis*
pseudomonas.biocyc.org	*Pseudomonas aeruginosa*
yeast.biocyc.org	*Saccharomyces cerevisiae*

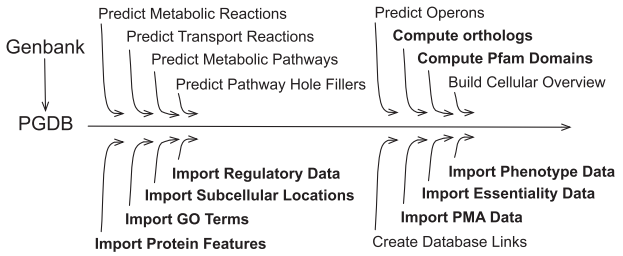

Figure 1. Processing steps involved in generating the BioCyc DBs. Recently added steps are shown in bold. No relative ordering is implied between the steps along the top and the steps along the bottom.

First, Pathway Tools converts the annotated genome from the Genbank format to its internal PGDB format. Next, the computational operations in the upper portion of Figure 1 are performed. Pathway Tools modules make the following predictions [2]. Metabolic and transport reactions and metabolic pathways are predicted [3] from the reactions and pathways in the MetaCyc DB [4]. Next occurs prediction of pathway hole fillers (genes that code for enzymes catalyzing reactions with no currently assigned enzyme) and prediction of operons using both structural and functional information [5].

Orthologs among BioCyc genomes are computed by software that runs large-scale bidirectional BLAST (version 2.2.23) comparisons among all pairs of proteins in the BioCyc genomes. We use a BLAST *E*-value cutoff of 0.001, with all other parameters at default settings. We define two proteins A and B as orthologs if protein A from proteome P_A and protein B from proteome P_B are bidirectional best BLAST hits of one another, meaning that protein B is the best BLAST hit of protein A within proteome P_B, and protein A is the best BLAST hit of protein B within proteome P_A. In rare cases, protein A might have multiple orthologs in proteome P_B, as explained below. The best hit of protein A in proteome P_B is defined by finding the minimal *E*-value among all hits in proteome P_B in the BLAST output, and collecting all the hits for A in proteome P_B that have the same minimal *E*-value. In other words, ties are possible, as in the case of exact gene duplications. We attempt to break ties using two methods: taking the hit with the maximum alignment length; and then taking the hit with the maximum alignment amino acid residue identity. For the first method, we compare the alignment lengths among all the hits of protein A in proteome P_B that share the same minimum *E*-value, and the protein in proteome P_B with the maximum alignment length is selected. For the second method, we compare the number of identical amino acid residues in the alignments between protein A and the hits of protein A in proteome P_B that share the same minimum *E*-value, and the protein in proteome P_B with the maximum number of identical amino acid residues is selected. In the case that ties still remain (as in the case of exact gene duplications), all ties are included in the final set of orthologs used by BioCyc. Thus, protein A could have multiple orthologs in P_B, such as if multiple proteins B1, B2, etc., exist in P_B, and have exactly the same regions align against protein A. BioCyc does not calculate paralogs.

Pfam [6] domains are identified in BioCyc proteins by running the Pfam software. Finally, zoomable cellular overview (metabolic map) diagrams are generated for each organism.

Next, data from several third-party DBs are imported into BioCyc, as shown in the lower portion of Figure 1. Protein-feature data, such as locations of enzyme active sites, phosphorylation sites and metal-ion binding sites, are loaded from UniProt [7],

as are Gene Ontology (GO) [8] annotations. Predicted subcellular localizations are loaded from PSORTdb [9]. Descriptions of promoters, transcription factor-binding sites and regulatory interactions are loaded from RegTransBase [10]. Organism phenotype data, such as aerobicity, are loaded from the National Center for Biotechnology Information (NCBI) BioSample DB, as are organism metadata, such as the geographical location of the site from which the sequenced organism was collected. Gene essentiality data have been loaded from the OGEE DB [11] and from individual articles. Phenotype microarray data have also been loaded from individual articles. We also generate Web links from BioCyc to other related DBs, such as UniProt, NCBI-Bioproject and BioSample.

BioCyc curation

After the preceding automated processing, some BioCyc DBs receive manual curation to integrate additional information and to remove some false-positive predictions. All in all, the information within the BioCyc DBs has been curated from 80 900 different publications, as shown in Table 2. The BioCyc DBs are organized into three tiers [12] to communicate the amount of manual curation that each DB has received:

- Tier 1 PGDBs have received at least one person-year of curation; some PGDBs have received person-decades of curation.
- Tier 2 PGDBs have received at least one person-month of curation.
- Tier 3 PGDBs have received no manual curation.

Some BioCyc PGDBs were contributed by groups outside SRI (for example, the *Chlamydomonas reinhardtii* PGDB was developed by the Carnegie Institution for Science, and the *Streptomyces coelicolor* PGDB was developed by the University of Warwick and the John Innes Centre). The authors of each PGDB are listed on the summary page that is displayed when a user changes the current PGDB.

Table 2. For those BioCyc version 21.0 PGDBs citing ≥100 references, we list the number of references cited by each PGDB (and from which the information in each PGDB was curated), sorted by number of citations

DB	Citations	Tier
MetaCyc	52 446	1
Escherichia coli K-12 substr. MG1655	31 555	1
Saccharomyces cerevisiae S288c	12 018	1
Bacillus subtilis subtilis 168	3682	2
Clostridioides difficile 630	2027	2
Mycobacterium tuberculosis H37Rv	1521	2
Chlamydomonas reinhardtii	1233	2
Candida albicans SC5314	623	2
Streptomyces coelicolor A3(2)	343	2
Synechococcus elongatus PCC 7942	284	2
Agrobacterium fabrum C58	257	2
Leishmania major strain Friedlin	212	1
Corynebacterium glutamicum ATCC 13032	184	2
Listeria monocytogenes 10403S	176	2
Candidatus Evansia muelleri	147	2

Note: In many cases, the curation was performed by BioCyc curators, and in other cases, the curation was performed by other DBs from which information was imported (e.g. from GO term curation or from UniProt protein-feature curation). For these PGDBs, we have removed from the citation counts those references shared with MetaCyc classes and metabolites (which were likely copied from MetaCyc during PGDB creation). MetaCyc and EcoCyc cite a number of common references because the EcoCyc pathway and enzyme data and their references are periodically copied from EcoCyc to MetaCyc.

The *Clostridioides difficile* 630 PGDB has undergone several recent curation enhancements. We updated its genome annotation from the recently revised RefSeq entry, and from the annotation from the MicroScope site [13]. We performed literature searches and curation updates for 213 proteins listed in MicroScope as having experimental evidence for their function in *C. difficile* or in the *Clostridioides* genus, as well as other genes encountered during the course of literature searches. Those proteins with experimental evidence in *C. difficile* are now annotated with experimental evidence codes and contain references to the literature from which their enhanced curation was derived.

Curation adds value to BioCyc PGDBs in many ways, and is a major factor in differentiating BioCyc from other bacterial genome PGDBs. All computational prediction methods make errors, including predictors of gene boundaries, protein function and metabolic pathways. Curators correct errors in those predictions, and they supplement computational predictions with information from the experimental literature. They also annotate experimentally known information with experimental evidence codes and literature citations to indicate high-confidence information.

Curators capture a wide variety of information in BioCyc PGDBs (Table 3) including protein functions, metabolic reactions and pathways and regulatory interactions of several types (such as allosteric regulation of enzymes, and control of gene expression via transcription factors and small RNAs).

Curators author mini-review summaries appearing in the protein, pathway and operon pages, which summarize findings from multiple publications and save users significant amounts of time in poring through the primary literature. For some BioCyc PGDBs, person-decades of curation work have been performed across tens of thousands of publications, resulting in large volumes of mini-review summaries, measured in textbook page equivalents: EcoCyc version 21.0 contains 2907 textbook-equivalent pages of summaries and MetaCyc version 21.0 contains 7897 such pages. Further, curators enter a wide range of experimentally determined information that cannot be inferred computationally, including enzyme activators and inhibitors, protein subunit structure, enzyme kinetic values, protein features (e.g. active site residues) and transcriptional regulatory interactions.

Although automated text mining software has shown gradual improvement over the years, its accuracy is still far from that of human curators. In addition, text mining systems are typically limited to extracting fewer types of data than the wide range of information that BioCyc curators capture. Perhaps most importantly, only human curators can correctly resolve the many disagreements, inconsistencies and errors found in the literature. Many metabolic pathways and enzymes are complex, and earlier reports often contain information that has been later partially or completely invalidated. For example, enzyme commission (EC) 2.3.1.111, mycocerosate synthase, was initially reported to release its product in the form of a coenzyme-A activated compound, but later, it was shown that the products remain bound to the enzyme at the end of catalysis because of a lack of a thioesterase function. A computer program reading through the conflicting reports would have great difficulty in reconciling the information from the different publications. An experienced human curator, on the other hand, can integrate the information and generate a review that consolidates all sources and provides an accurate review of current knowledge.

Expansion of bioinformatics tools

The BioCyc.org Web site offers, to our knowledge, the most extensive set of bioinformatics tools of any microbial genome

Table 3. Datatypes available in PGDBs, and statistics on the number of objects of each type in various PGDBs

DB tier	*Escherichia coli* K-12 substr. MG1655 1	*Bacillus subtilis* subtilis 168 2	*Synechococcus elongatus* PCC 7942 2	*Mycobacterium tuberculosis* Beijing/NITR203 3
Data type				
Genome metadata	2	2	4	0
Genes	4657	4440	2719	4206
Operons	3564	1604	1982	2680
Promoters	3850	1193	44	0
Transcription factor-binding sites	2918	763	36	0
Terminators	303	1146	0	0
Proteins	5719	4407	2832	4127
Protein features	4223	3029	0	0
Gene Ontology terms	5733	3927	2518	4
Metabolites	2758	942	990	1169
Metabolic reactions	1712	1158	1100	1450
Metabolic pathways	396	269	230	285
Transport reactions	1526	1048	953	1148
Genetic regulatory networks	3438	788	41	0
Evidence codes	134 561	58 658	15 098	3137
Growth media	436	1	2	0
Gene essentiality	4239	4217	2421	0

Note: Different DBs contain different proportions of these datatypes depending on factors such as the amounts of data available in DBs from which BioCyc imports information, and the amount of data curated from the literature. Typically, DBs that have received more curation will have objects of a wider range of datatypes.

portal (Table 4). Many of the tools provide visualization services that aid the user in navigating the large and complex information space within BioCyc.

This section surveys a number of recent developments in the BioCyc software tools. For a comprehensive description of the Pathway Tools software behind BioCyc, see [1].

Run Metabolic Model

The 'Run Metabolic Model' command allows users to solve steady-state metabolic models based on flux balance analysis [14]. Metabolic models are generated from the reactions stored in a PGDB by the MetaFlux component of Pathway Tools [1]. Users must login to their BioCyc account to be able to use the command 'Run Metabolic Model'. Existing public models, or a user's own private models, can be executed. For example, by selecting the *Escherichia coli* K-12 substr. MG1655 organism, the 'Run Metabolic Model' command will open up a new Web page and show a list of several public models available for that organism (by the owner 'BRG SRI') and any metabolic models that the user has created, some of which are probably private.

Click the 'Select' button of a model to analyze and execute that model. After clicking the 'Execute' button, the 'Results' tab will provide the biomass flux of that model, three buttons ('Show Solution File', 'Show Log File' and 'Show Fluxes on Cellular Overview') to further analyze the solution and the list of all reactions that are active (i.e. reactions with nonzero fluxes) in the model. The solution file shows much more detail about the solutions, such as the fluxes of all biomass metabolites, nutrients and secreted metabolites. The log file gives a complete list of all reactions that are in the model, including the non-active reactions, the reactions that are blocked; the instantiated reactions; and more. The 'Show Fluxes on Cellular Overview' button, if clicked, will open the Cellular Overview metabolic map diagram of the organism with the reactions and pathways highlighted according to the fluxes of the reactions. The model specification can be seen by selecting the tabs 'Nutrients', 'Reactions', 'Biomass' and 'Secreted Metabolites'. From the Web page displaying one model, you can go back to the list of models by clicking 'View Models' near the top left corner of the Web page.

The publicly available models can be copied and modified. To copy a model, click the 'Copy' button and enter a new name for it. The copy is private in your own account and can be modified and solved at will. For example, the nutrients and secreted metabolites of the model can be modified to execute under different growth conditions (e.g. anaerobic). If desired, by clicking the 'Make Public' button, the model can be shared with all the users that have access to the Web server.

You can learn more about the Run Metabolic Model tool via the 'Getting Started Guide' link on the Web page listing the models available.

New search tools

As the number of organisms in BioCyc has grown, we have introduced new search tools to help users find organisms of interest. Clicking 'change organism database' in the upper right corner of the Web site brings up the organism-selector dialog, which enables a user to search for organisms by name, by organism taxonomy and by phenotypic and metadata properties. The name-based search finds any prefix of the genus, species or strain name. The taxonomy search uses a hierarchical browser of the NCBI taxonomy DB. For the phenotypic/metadata search, the user first selects a property (such as 'biotic relationship') and then selects the value of interest for that property (such as whether the desired organism is free living, parasitic or symbiotic). Additional available properties include whether the organism is a pathogen of humans, animals or plants; the human microbiome body site from which the organism was collected; and the depth or altitude at which the organism sample was collected. Metadata searches include number of GO terms annotated within the DB and number of regulatory interactions within the DB. Multiple properties can be queried at once (combined using AND or OR) by clicking the 'Add Constraint' button.

Table 4. BioCyc software tools

Genome tools	
Gene/Protein/RNA Page	Presents information about individual genes and their products
Genome Browser	View genes and other genome regions at variable magnification, from full genome on one screen to sequence level
Comparative Genome Browser	Align genome regions from multiple organisms around shared orthologous genes
Genome Poster	Printable poster containing full-genome diagram
Gene Ontology Browser	Hierarchical browser for navigating within GO
Sequence Alignment Viewer	Align nucleotide or amino acid sequences
BLAST Sequence Search	Search nucleotide or amino -acid sequences against individual BioCyc genomes or against all BioCyc genomes
Sequence Pattern Search	Search short nucleotide or amino acid sequences including wild cards against a BioCyc genome
Gene Regulatory Network Browser	Visualize and navigate within complete organism regulatory network
Operon Page	Presents information about an operon and its regulatory sites and regulators
Comparative Analysis	Compare genome and pathway information across organisms
Metabolism tools	
Pathway Page	Presents information about individual metabolic pathways
Pathway Collage Diagrams	Personalized multi-pathway diagrams: the user chooses a set of pathways, positions them relative to one another and defines connections among them
Metabolic Network Browser	Zoomable browsing of organism-specific metabolic network diagrams
Metabolic Network Posters	Printable organism-specific metabolic network diagrams
Run Metabolic Models	Execute quantitative steady-state metabolic models
Chokepoint Analysis	Compute potential antimicrobial drug targets based on metabolic network choke points
Dead-End Metabolite Analysis	Find metabolites that are not producible or not consumable
Metabolic Route Search	Search for optimal paths through reaction network connecting starting and ending metabolites
Omics Data Analysis	
Paint Data on Metabolic Network	Color full zoomable metabolic network diagram with omics data
Paint Data on Pathway Diagram	Color individual pathway diagrams with omics data
Paint Data on Pathway Collage	Color pathway collage diagrams with omics data
Paint Data on Genome Browser	Color single-screen genome diagram with omics data
Paint Data on Regulatory Network	Color full regulatory network diagram with omics data
Enrichment Analysis	Compute statistical enrichment of GO terms, pathways, regulators
Other tools	
Update Notifications	Receive notification of curation updates in declared research interest areas
Advanced Search	Create SQL-like complex DB searches
Cross-Organism Search	Perform text searches across all of BioCyc or specified groups of organisms
Multi-Organism Route Search	Search for metabolic routes that cross-multiple organisms such as the gut microbiome
SmartTables	Define groups of genes, pathways, metabolites, etc., and manipulate those groups as a programmer would

To facilitate comparative analyses, a new multi-organism search tool is available under Search → Cross Organism Search. The user can specify what set of organisms (DBs) to search in several alternative ways, such as by specifying taxonomic groups (e.g. 'Archaea' or 'Coriobacteriia'), by specifying organisms by names, or by selecting organisms according to their phenotypic properties (e.g. selecting all symbionts). A user can also save lists of organisms for later use within SmartTables.

A cross-organism search enables the user to search a designated set of organisms for search terms in specific object types. The user specifies the types of objects to search for (e.g. genes or metabolites), and one or more search terms (e.g. 'trpA' or 'acetaldehyde'). The tool returns a table indicating what objects from what organisms matched the requested search.

Redesigned gene/protein pages and metabolite pages

We have redesigned BioCyc gene/protein pages to modernize their look and feel and to make it easier for scientists to find the information they seek. The new design provides a summary of commonly used information at the top, with additional information available via the tabs just below the table. For example, the 'Protein Features' tab depicts protein features such as metal-ion binding sites and enzyme active sites; the 'Operons' tab depicts the operon(s) containing the gene. The 'Show All' tab combines information from all tabs into one page, which is convenient when searching the Web page for terms of interest.

A new menu, called the right-sidebar menu, is available along the right side of gene/protein pages and most other BioCyc pages. Its content varies depending on the page type currently displayed (e.g. different operations are available for gene pages versus metabolite pages). Operations available at gene pages include retrieving the nucleotide and amino acid sequences for the gene/protein, and retrieving arbitrary nucleotide sequences surrounding a gene, or for any region of the genome. Other gene-page operations include creating a multi-genome alignment (using the Pathway Tools comparative genome browser) and a multiple sequence alignment [computed using MUSCLE [15] and displayed using the Sol Genomics Network alignment viewer (https://sgn.cornell.edu/tools/align_viewer/index.pl)] for the current gene and specified orthologs.

Metabolite pages have been redesigned along similar lines. They contain a table at the top that summarizes important information, along with tabs to select additional information such as the reactions in which a metabolite occurs.

SmartTables

SmartTables [16] enable scientists to define and store lists of objects from any BioCyc DB, such as lists of genes, proteins,

ADD TRANSFORM COLUMN: choose a transform... | ADD PROPERTY COLUMN: choose a property | ENRICHMENTS: choose an enrichment...

1 2 3 4 5 6 7 8 9 10 11 12 13 14 Next Show all

	(All-Genes STRING)	Product	Pathways of gene
1	accA	acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha	biotin-carboxyl carrier protein assembly
2	accB	acetyl-CoA carboxylase subunit	biotin-carboxyl carrier protein assembly
3	accC	acetyl-CoA carboxylase subunit	biotin-carboxyl carrier protein assembly
4	accD	acetyl-CoA carboxylase carboxyl transferase subunit beta	biotin-carboxyl carrier protein assembly
5	acpP	acyl carrier protein	
6	acpS	holo-ACP synthase	acyl carrier protein metabolism II (mammalian)
7	addB	ATP-dependent helicase/nuclease subunit B	
8	adk	adenylate kinase	superpathway of purine nucleotides *de novo* biosynthesis I superpathway of adenosine nucleotides *de novo* biosynthesis II superpathway of adenosine nucleotides *de novo* biosynthesis I superpathway of purine nucleotides *de novo* biosynthesis II adenosine ribonucleotides *de novo* biosynthesis
9	alaS	alanine--tRNA ligase	tRNA charging
10	argS	arginyl-tRNA ligase	tRNA charging
11	asd	aspartate-semialdehyde dehydrogenase	superpathway of *S*-adenosyl-L-methionine biosynthesis superpathway of L-threonine biosynthesis superpathway of L-isoleucine biosynthesis I L-homoserine and L-methionine biosynthesis superpathway of L-methionine biosynthesis (transsulfuration) L-lysine biosynthesis II L-homoserine biosynthesis
12	asnC	asparagine--tRNA ligase	tRNA charging

Figure 2. SmartTable showing essential genes and their metabolic pathways in *C. difficile* 630.

metabolites, pathways or sequence regions (e.g. SNPs). Using SmartTables, a scientist can browse and explore a group of objects. They can transform a group of objects to a set of related objects (such as transforming a metabolite set to the set of pathways those metabolites are involved in). Users can also perform analyses such as statistical enrichment analysis (e.g. to understand what functional categories are shared by the differentially regulated genes from a transcriptomics experiment). Scientists can share SmartTables with specific colleagues or with the public, and can use them to supplement a publication by providing online gene or metabolite sets.

Users must create a BioCyc account to create SmartTables. SmartTables can be created to contain results from different types of BioCyc query operations (look for the button 'Turn into a SmartTable'). They can also be created from a file—we defined the public SmartTable at https://biocyc.org/group?id=biocyc14-1553-3655492599 by uploading a file listing the essential genes determined for *C. difficile* R20291 by Dembek *et al.* [17], and then adding the orthologous genes and gene products in *Bacillus subtilis* and *E. coli* as additional columns.

We will use this SmartTable to illustrate some general capabilities of SmartTables by investigating the question of which metabolic pathways these essential genes are involved in. We begin by creating a separate SmartTable containing the orthologs from strain 630 of this essential gene set, by clicking the '+' at the top of the column labeled 'Gene/Locus-Ids in Strain 630'. Next, from the 'Add Property Column' menu directly above the SmartTable, select 'Product' to add a new column containing the gene products, and from the 'Add Transform Column' menu directly above the SmartTable, select 'Pathways of Gene' to add a column listing the metabolic pathway(s) (if any) in which these gene products participate; the result is shown in Figure 2.

To see these same data from a different perspective, click the '+' above the 'Pathways of Gene' (third) column, which will create a new SmartTable listing each metabolic pathway, and the essential genes within that pathway. Another way to see the data from the SmartTable in Figure 2 is to run the operation 'Paint Data → on Cellular Overview' from the right-sidebar menu (be sure the first column in the SmartTable is selected, by clicking on it). This operation will display the set of genes within the SmartTable on a zoomable metabolic map diagram for strain *C. difficile* 630.

Among the other operations provided for SmartTables are adding and deleting rows individually, using a filtering operation to remove rows that meet criteria such as containing a search string, and performing set operations such as union and intersection between two SmartTables. SmartTables also offer views of the nucleotide and amino acid sequences of genes and proteins, and of the chemical structures of metabolites.

Pathway collages

For many years, BioCyc has provided the ability for users to customize its images of metabolic pathways. The command 'Customize or Overlay Omics Data on Pathway Diagram' from the right-sidebar menu of any pathway page enables users to control which elements of the pathway diagram are visible (gene names, EC numbers, etc.), and to overlay gene expression, metabolomics or reaction-flux data on the pathway.

Pathway collages are a new way of creating diagrams depicting interactions among several metabolic pathways, and were suggested by Prof. Tricia Kiley of the University of Wisconsin. Define a SmartTable containing the pathways to include in the pathway collage (such as by creating a new SmartTable and then adding the pathways by name). Then use the right-sidebar menu command 'Export → Export Pathways to Pathway Collage' to create the pathway collage within a Web browser. The commands available within the pathway collage builder include dragging

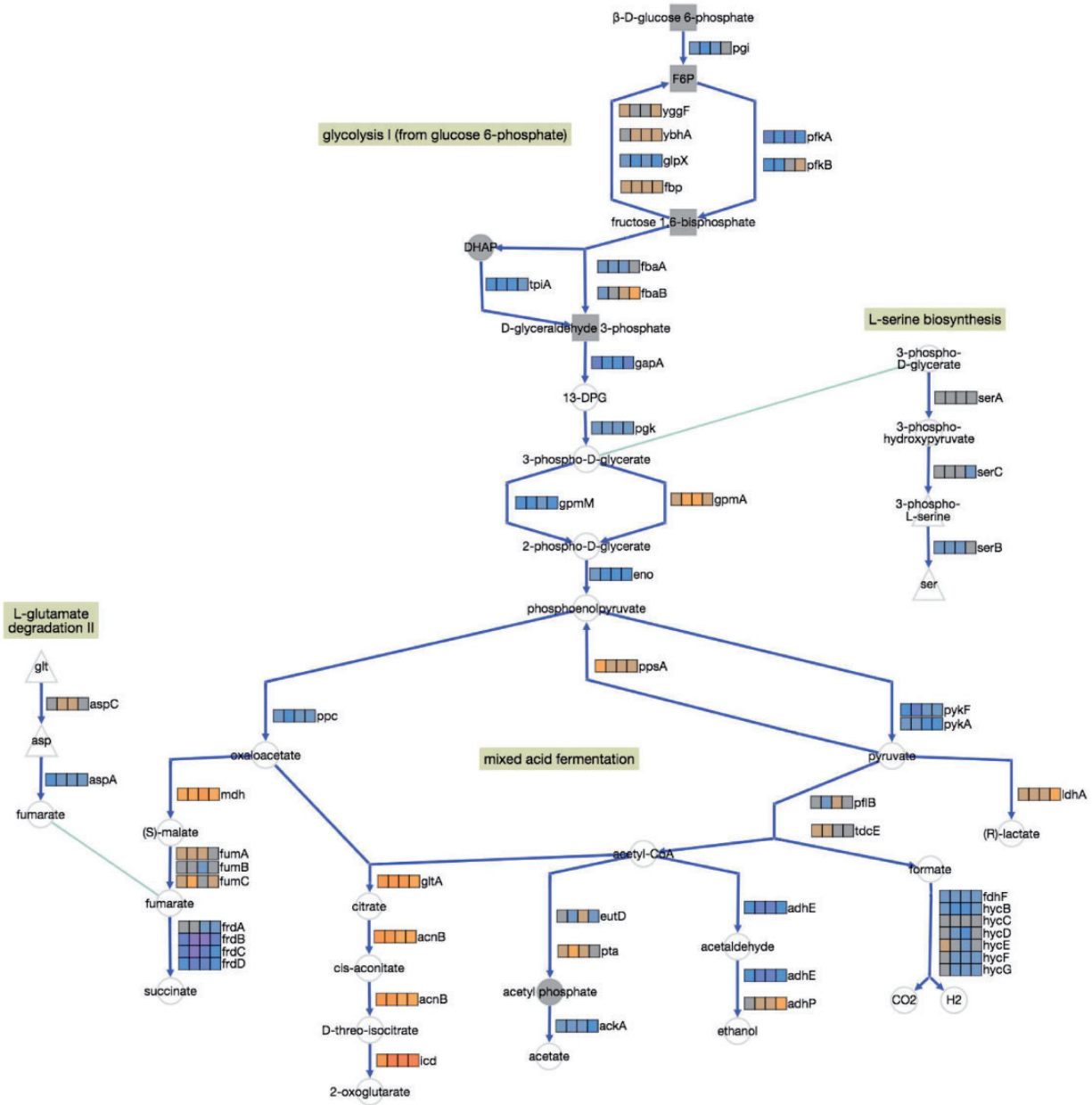

Figure 3. *Escherichia coli* gene expression data from an anaerobic to aerobic transition (Gene expression omnibus id GDS2364) superimposed on a pathway collage. Each horizontal row of squares indicates the expression levels of one gene across the four oxygen levels depicted, where the leftmost square is anaerobic and the rightmost square is the highest oxygen concentration. Blue and purple indicate low expression, gray indicates intermediate expression and orange indicates high expression.

pathways to new positions, creating connection lines between metabolites, changing the visual appearance of gene and metabolite names and adding omics data to the diagram. An example collage of *E. coli* pathways is shown in Figure 3. Collages can be exported to PNG files for use in publications.

Pan-genome DBs

We are introducing pan-genome PGDBs in BioCyc that integrate gene and pathway information from a large number of sequenced strains into one DB. Pan-genome PGDBs illuminate the set of gene families found across the species. Pan-genome PGDBs now exist for *Listeria monocytogenes*, *Mycobacterium tuberculosis*, *C. difficile* and *Pseudomonas aeruginosa*, and can be found by searching for the phrase 'pan-genome' within the organism selector.

The following steps are taken to construct a Pan-Genome PGDB for species S:

- Create an empty PGDB for S.
- Choose a set of PGDBs for strains of S for which computed orthologs are available.
- Choose a so-called 'lead PGDB' from the preceding set of available strain-specific PGDBs. For example, we chose *M. tuberculosis* H37Rv as the lead PGDB for species *M. tuberculosis* because of its status as a highly studied strain.
- Import the lead PGDB's replicons, genes, proteins, reactions and pathways into the pan-genome PGDB.

- Visit every other strain PGDB from the chosen set. For each protein-coding gene in that PGDB, check whether it is an ortholog of any gene already residing in the pan-genome PGDB. If so, record the existence of the ortholog in the gene in the pan-genome PGDB.

If no ortholog was found, then import the new gene from the other strain PGDB, along with its proteins and any reactions and pathways that are not yet in the pan-genome PGDB. Finally, add the nucleotide sequence of the newly added gene to an 'artificial replicon', which accumulates all these other genes (separated by spacers consisting of several N nucleotides).

The new gene will thereafter also be checked for orthology in future comparison rounds with additional strain PGDBs. The end result will be that many genes, both on the replicons from the lead PGDB and on the 'artificial replicon', will have orthologs recorded, and some genes from the lead PGDB and the other strain PGDBs will be unique and have no orthologs at all.

When viewing the Cellular Overview for a pan-genome PGDB, two special highlighting commands are made available. Highlighting the core genes shows all the reactions of the genes that are shared among all the strain PGDBs; in other words, each gene has orthologs to all the other strains. Highlighting the unique genes shows all the reactions of the genes that have no orthologs at all, and are thus contributed by one strain.

RouteSearch, atom mappings and Gibbs free energies

BioCyc reaction pages depict atom mappings for most reactions. The atom mapping of a reaction identifies for each reactant non-hydrogen atom its corresponding atom in a product compound. For a given reaction in a BioCyc PGDB, atom mapping data are obtained from the same reaction (reaction having the same reaction identifier) in MetaCyc. We computed MetaCyc atom mappings using an algorithm that minimizes the overall cost of bonds broken and made in the reaction, given assigned propensities for bond creation and breakage [18]. Of the 14 051 reactions in MetaCyc, 12 356 (87.9%) have computed atom mappings. Our analysis [18] has found a low rate of errors (<3%) in our computed atom mappings.

RouteSearch (see Metabolism → Metabolic Route Search) [19] is a software tool for finding routes in the metabolic reaction network of an organism. Given a starting compound, a target compound and other parameters, the tool finds the best (least cost) routes between these compounds by taking into account atom conservation (routes that conserve more atoms from the starting compound are considered better), reaction path length and adding a minimum number of foreign reactions from MetaCyc. RouteSearch uses the precomputed atom mappings of the reactions involved in the routes to calculate the number of conserved atoms.

Gibbs free energies are provided for a large number of reactions and compounds in BioCyc, based on data in the MetaCyc DB. We calculated standard Δ Gibbs free energies for reactions and compounds in MetaCyc, that is $\Delta_r G'^{\circ}$ and $\Delta_f G'^{\circ}$, at pH 7.3 and ionic strength 0.25.

Computational access to BioCyc data

A variety of REST-based Web services offer programmatic access to the BioCyc data via HTTP GET or POST requests [20]. A set of defined queries enables retrieval of data for a single object (such as a gene or a reaction) or collection of related objects (such as all the genes in a pathway) in XML format [21]. More complex Web service queries to BioCyc, of power on the order of SQL, can be constructed using the powerful BioVelo Query Language [22]. Web services also provide access to pathway data in BioPAX [23] format. Additional Web services enable mapping of identifiers from external DBs, and retrieval of metabolites by chemical formula, InChI key and/or monoisotopic molecular weight. A variety of visualization services and SmartTable manipulation operations provide access to advanced BioCyc capabilities, and are further described at [20].

BioCyc data are available for bulk download in several different file formats [24]. In addition to tab-delimited tables and our own internal attribute-value format [25], subsets of the data are made available in SBML [26], BioPAX [27], GO [28], GenBank [29] and FASTA [30] formats.

Users who install the Pathway Tools software locally can access and update data directly via our application programming interfaces (APIs), available for Python, R, Java, Perl and Common Lisp. The PythonCyc [31], RCyc [32], JavaCyc [33] and PerlCyc [34] packages, which provide API access to their respective languages, must be downloaded and installed separately from the main Pathway Tools distribution.

BioCyc subscription model

Model-organism DBs such as EcoCyc, *Saccharomyces* Genome DB, FlyBase, Mouse Genome DB and Rat Genome DB see high usage rates. Thus, it is fairly clear that curated genome DBs are a critical part of the scientific information infrastructure for sequenced organisms that are studied by large scientific communities and that have important applications (e.g. *M. tuberculosis*, which is a significant pathogen, and *B. subtilis*, which sees widespread use in biotechnology).

It has also become clear that the cost of DB curation is fairly modest and can attain low error rates. For example, the cost of curation for the EcoCyc DB was $219 per curated article over a 5 year period, which is modest when compared with the costs of the projects that generated the research to be curated: for EcoCyc, we estimated that curation cost to be ∼0.088% of the cost of the research projects that generated the research and to be 6–15% of the cost of open-access publication fees for publishing the curated research [35]. The EcoCyc error rate was measured to be 1.40% [36].

Despite the fact that a number of bioinformatics groups have put forward the preceding arguments over a 15 year period, government funding agencies have not provided funds for additional needed DB curation projects, particularly for bacteria. Thus, in 2016, we decided to convert BioCyc to a subscription model to raise revenue for the curation of BioCyc DBs. Subscriptions to BioCyc are available to individuals and to institutions such as companies and university libraries. Subscription costs are similar to the costs of journal subscriptions, and depend on usage level. Subscription revenues are invested in a nonprofit basis in BioCyc curation, operation, sales and marketing. Access to EcoCyc and MetaCyc DBs remains free because these DBs are still supported by government grants.

Conclusions

We have outlined some of the recent improvements to BioCyc. Additional improvements to the Pathway Tools software are described in a recent article [2]. The data content and software tools within BioCyc will continue to evolve. The human microbiome and metabolomics data analysis are two major topics of our current grant period.

How to learn more

A number of online information sources are available for BioCyc including online instructional videos [37], a how-to guide for the BioCyc Web site [38], a guide to the concepts and methods behind BioCyc [39] and a guide to the data content of BioCyc [40].

To receive monthly updates and explanations regarding new developments in BioCyc, please subscribe to the BioCyc mailing list by sending an e-mail to biocyc-users-request@ai.sri.com with the word 'subscribe' in the subject.

Key Points

- BioCyc.org is a microbial genome Web portal that combines sequenced genomes, computationally inferred data and curated information from the scientific literature.
- BioCyc provides an extensive range of query tools, visualization services and analysis software.
- BioCyc SmartTables is a unique tool that enables biologists to perform analyses that previously would have required a programmer's assistance, such as performing programmatic transformations on sets of objects.

Funding

The National Institute of General Medical Sciences of the National Institutes of Health (grant numbers R01GM080746, R01GM75742 and R01GM077678). 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

1. Karp P, Latendresse DM, Paley SM, *et al*. Pathway Tools version 19.0: software for pathway/genome informatics and systems biology. *Brief Bioinform* 2016;**17**:877–90.
2. Karp PD, Latendresse M, Paley SM, *et al*. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform* 2015. doi:10.1093/bib/bbv079.
3. Karp PD, Latendresse M, Caspi R. The pathway tools pathway prediction algorithm. *Stand Genomic Sci* Dec 2011;**5**(3):424–9.
4. Caspi R, Billington R, Ferrer L, *et al*. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* 2016;**44**(D1): D471–80.
5. Romero P, Karp PD. Using functional and organizational information to improve genome-wide computational prediction of transcription units on Pathway/Genome Databases. *Bioinformatics* 2004;**20**:709–17.
6. Finn RD, Coggill P, Eberhardt RY, *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;**44**(D1):D279–85.
7. UniProt Consortium. Update on activities at the universal protein resource (UniProt) in 2013. *Nucleic Acids Res* 2013;**41**:D43–7.
8. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 2015;**43**:D1049–56.
9. Peabody MA, Laird M, Vlasschaert RC, *et al*. PSORTdb: expanding the bacteria and archaea protein subcellular localization database to better reflect diversity in cell envelope structures. *Nucleic Acids Res* 2016;**44**(D1):D663–8.
10. Cipriano MJ, Novichkov PN, Kazakov AE, *et al*. RegTransBase–a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes. *BMC Genomics* 2013;**14**:213.
11. Chen R, Mias GI, Li-Pook-Than J, *et al*. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;**148**(6):1293–307.
12. List of BioCyc Pathway/Genome Databases. https://biocyc.org/biocyc-pgdb-list.shtml.
13. MicroScope Home Page. https://www.genoscope.cns.fr/agc/microscope/home/index.php.
14. Latendresse M, Krummenacker M, Trupp M, *et al*. Construction and completion of flux balance models from pathway databases. *Bioinformatics* 2012;**28**:388–96.
15. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**(5): 1792–7.
16. Travers M, Paley SM, Shrager J, *et al*. Groups: knowledge spreadsheets for symbolic biocomputing. *Database* 2013.
17. Dembek M, Barquist L, Boinett CJ, *et al*. High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. *mBio* 2015;**6**(2):e02383.
18. Latendresse M, Malerich J, Travers PM, *et al*. Accurate atom-mapping computation for biochemical reactions. *J Chem Inf Model* 2012;**52**:2970–82.
19. Latendresse M, Krummenacker M, Karp PD. Optimal metabolic route search based on atom mappings. *Bioinformatics* 2014;**30**:2043–50.
20. Pathway Tools Web Services. https://biocyc.org/web-services.shtml.
21. Guide to ptools-xml. https://biocyc.org/ptools-xml-guide.shtml.
22. The BioVelo Query Language. https://biocyc.org/biovelo Language.html.
23. Demir E, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**(12):935–42.
24. BioCyc and Pathway Tools Download Information. https://biocyc.org/download.shtml.
25. Pathway Tools Data-File Formats. http://bioinformatics.ai.sri.com/ptools/flatfile-format.html.
26. SBML. http://www.sbml.org/.
27. BioPAX. http://www.biopax.org/.
28. GO Annotation File Formats. http://geneontology.org/page/go-annotation-file-formats.
29. Genbank Format. http://www.ncbi.nlm.nih.gov/collab/FT/#7.1.2.
30. FASTA Format. http://www.ncbi.nlm.nih.gov/blast/fasta.shtml.
31. PythonCyc API to Pathway Tools. http://bioinformatics.ai.sri.com/ptools/pythoncyc.html.
32. RCyc API to Pathway Tools. https://github.com/taltman/RCyc/blob/master/DESCRIPTION.
33. JavaCyc API to Pathway Tools. http://solgenomics.net/downloads/index.pl.
34. PerlCyc API to Pathway Tools. http://solgenomics.net/downloads/perlcyc.pl.
35. Karp PD. How much does curation cost? *Database* 2016.
36. Keseler IM, Skrzypek M, Weerasinghe D, *et al*. Curation accuracy of model organism databases. *Database* 2014:1–6.
37. BioCyc Webinars. https://biocyc.org/webinar.shtml.
38. How to Use a Pathway Tools Website. https://biocyc.org/PToolsWebsiteHowto.shtml.
39. Pathway/Genome Database Concepts Guide. https://biocyc.org/PGDBConceptsGuide.shtml.
40. BioCyc Database Guide. https://biocyc.org/BioCycUserGuide.shtml.