

# BioCyc Microbial Genomes and Metabolic Pathways Web Portal

The BioCyc web portal from SRI International contains genome and metabolic-pathway information for 9,300 microbes. BioCyc databases are unique in integrating a diverse range of data and providing a high level of curation for important microbes. BioCyc curators summarize and synthesize information from thousands of publications, saving scientists time in literature research, and integrating data for large-scale computational analyses.

BioCyc Pathway/Genome Databases (PGDB) describe the genome of an organism, as well as its biochemical pathways and (for a small fraction of organisms) its regulatory network. New expanded releases occur three times per year, often with new databases present. Two members of the BioCyc collection, the EcoCyc [1] and MetaCyc [2] databases, are derived from more than three decades of literature-based curation of genome and pathway data.

The HumanCyc database provides a curated collection of many human metabolic pathways and enzymes.

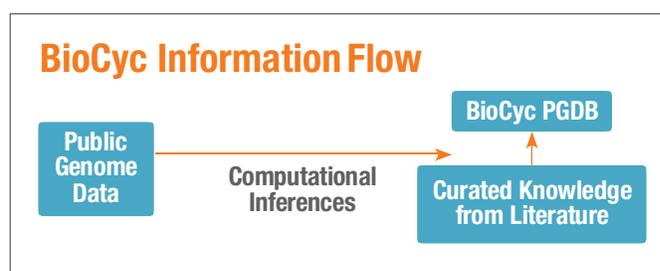
BioCyc bioinformatics tools combine unparalleled breadth and user friendliness. BioCyc provides a unique set of visualization tools to speed comprehension of its extensive and complex data.

## BIOCYC APPLICATIONS

BioCyc databases integrate extensive data for each organism, and provide platforms for analysis of large-scale datasets.

BioCyc enables scientists to pursue several use cases:

- BioCyc is a massive encyclopedic reference on microbial genes, metabolites, and pathways that integrates information from many sources. Scientists consult BioCyc to save large amounts of time finding, understanding, and synthesizing material from the primary literature.
- BioCyc is a genome informatics and comparative genomics platform
- BioCyc enables exploration of a vast set of biological networks
- BioCyc provides gene-expression, metabolomics, and multi-omics analysis tools
- BioCyc provides executable metabolic models for a small but growing set of organisms



## THE BIOCYC DATABASE COLLECTION

Each PGDB in the BioCyc collection describes the genome and metabolic network of a single organism.

All BioCyc PGDBs include:

- The organism's annotated genome
- Predicted metabolic pathways
- Predicted pathway hole fillers (genes coding for missing enzymes in metabolic pathways)
- Bacterial PGDBs include predicted operons
- Predicted reaction atom mappings (enable users to trace atoms from reactants to products)
- Predicted Gibbs free energies of formation
- Generated metabolic network poster and genome poster

The exact types of data present in different databases will vary. Many databases also include:

- Protein subcellular locations, enzyme kinetics data, protein features, Gene Ontology terms, predicted Pfam domains
- Curated regulatory information including promoters, operons, transcription-factor binding sites
- Ortholog relationships to other BioCyc genomes
- Organism phenotype data
- Database links

## BIOCYC WEBSITE AND DATABASE COLLECTION

- |  |   |
|--|---|
| <ul style="list-style-type: none"> <li>• Microbial genomes web portal for 9,300 organisms</li> <li>• Curated genome and metabolic pathway data derived from 66,000 publications</li> </ul> | <ul style="list-style-type: none"> <li>• Extensive genome informatics and comparative genomics tools</li> <li>• Omics data analysis tools</li> <li>• Quantitative metabolic models</li> </ul> |
|--|---|

BioCyc databases are organized into three tiers to reflect their quality levels:

**Tier 1 Pathway/Genome Databases** have received at least one person-year, and in some cases, person-decades, of manual curation:

- EcoCyc: The data in this *Escherichia coli* K-12 MG1655 database have been gathered during two decades of literature-based curation from more than 32,000 articles [1].
- EcoCyc describes the metabolic, transport, and regulatory machinery of *E. coli*. EcoCyc curators have authored the equivalent of 2,800 textbook-pages of mini-review summaries for *E. coli* genes and pathways.
- MetaCyc: Contains 2,500 metabolic pathways and 14,000 biochemical reactions from all domains of life. MetaCyc data and commentary were gathered from 51,000 publications to provide a comprehensive metabolic encyclopedia whose mini-review summaries encompass the equivalent of 7,700 textbook pages[2].
- YeastCyc: This highly curated metabolic network for *Saccharomyces cerevisiae* was curated from 2,600 publications.
- HumanCyc: This database was derived from a computational pathway analysis of the human genome, followed by literature-based curation of human pathways and enzymes.

**Tier 2 Pathway/Genome Databases** were computationally generated, and then received significant subsequent curation. BioCyc Tier 2 databases include

- BsubCyc: The curated metabolic and regulatory networks for this database cite 3,600 publications. 160 regulatory genes control 1,100 regulated genes.

- *Mycobacterium tuberculosis*, *Corynebacterium glutamicum*, *Peptoclostridium difficile*, *Bacillus anthracis*, *Francisella tularensis*, *Helicobacter pylori*, *Vibrio cholerae*, and others.

**Tier 3 Pathway/Genome Databases** were computationally generated with no subsequent curation.

## BIOCYC INFORMATICS TOOLS

BioCyc provides a powerful and comprehensive set of features for querying, visualization, and analysis [3]. These tools help scientists find and digest information quickly.

## GENOME INFORMATICS TOOLS

- Search for genomes by name, taxonomy, phenotypic properties.
- Gene information page
  - Retrieve amino-acid sequence and nucleotide sequence of arbitrary genome region.
  - Query genes by gene name, accession number, sequence length, replicon position, protein properties (pI, MW, protein features, subcellular location, ligand), GO terms.
- Transcription-unit information page.
- Genome Browser (Figure 4) depicts genomic regions at user-selected resolution with semantic zooming that reveals new features at higher resolutions. Visible features include pseudogenes, promoters, transcription-factor binding sites, repeats, terminators, nucleotide sequence. Zoom to sequence. Generate genome poster.
- BLAST search sequence-pattern search via patmatchMap SNPs to genes and show effects on translation.

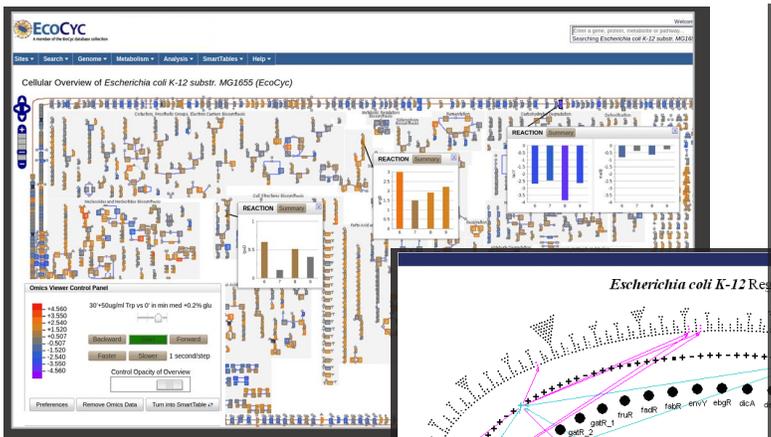


Figure 1: The Cellular Omics viewer paints omics datasets onto a diagram of the cellular biochemical network. Reaction lines can be colored with gene expression, proteomics, or reaction flux data; compound nodes can be colored with metabolomics data. Multi-omics data can be analyzed by coloring data onto reactions and metabolites simultaneously. Omics pop-ups graph omics data values using bar graphs, heat maps, or X-Y plots.

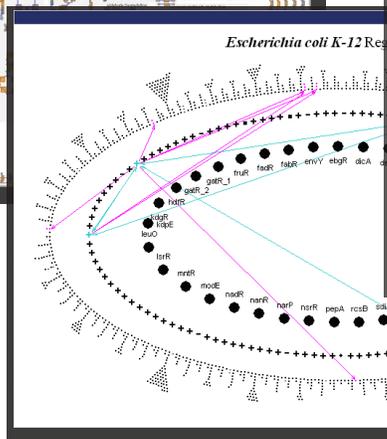


Figure 2: The Regulatory Overview depicts the transcriptional regulatory network in a PGDB. Here, the *E. coli* regulatory network is highlighted to show genes that regulate the *gntR* gene (blue lines), and the genes that are regulated by *gntR* (purple lines). The inner two rings are populated by transcription factors and sigma factors; the outer ring contains other genes.

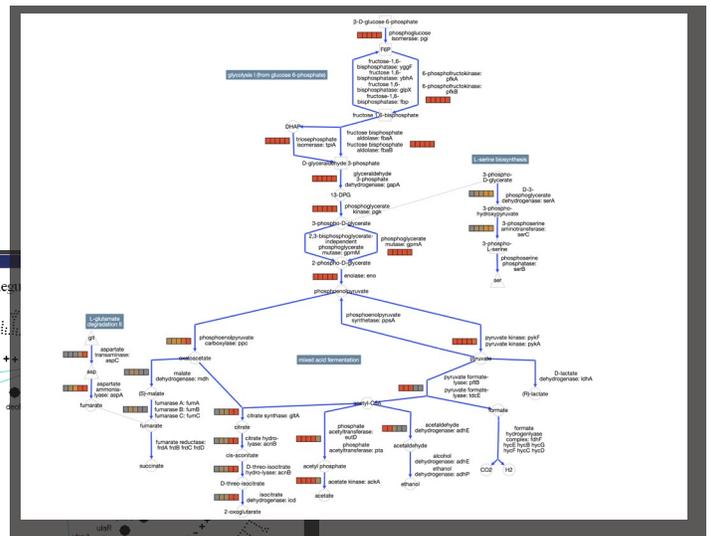
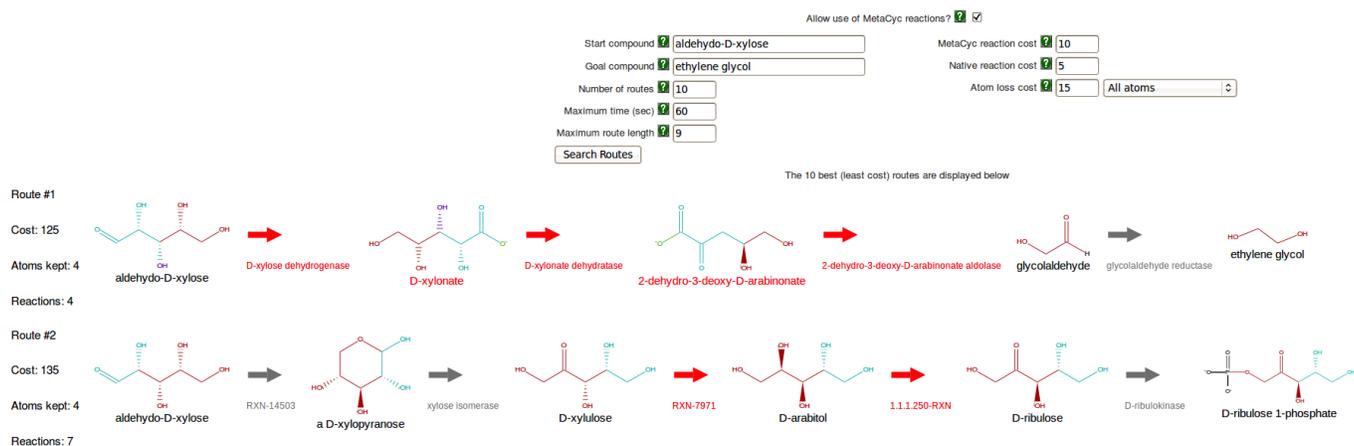


Figure 3: Pathway collages depict a user-selected set of metabolic pathways.





## COMPARATIVE GENOMICS TOOLS

- Comparative genome browser (Figure 4, Right) aligns chromosomal regions from multiple genomes at orthologous genes
- Sequence alignments
- Compare pathway, reaction, metabolite, and protein complements of specified organisms
- Quick navigation between corresponding entities (e.g., genes, pathways, metabolites) in different organisms.
- Cross-organism search finds genes, metabolites, pathways across BioCyc organisms

## WEB EXECUTION OF METABOLIC MODELS

- Execute steady-state metabolic flux models using flux-balance analysis
- Specify organism nutrients, secreted metabolites, and biomass metabolites synthesized by metabolic network
- Predict reaction flux rates, cellular growth rates
- Visualize reaction fluxes on metabolic network diagram
- Initial models available for *Escherichia coli*, *Eubacterium rectale*, *Bacteroides thetaiotamicron*; additional organisms to come

## ADVANCED DATABASE ACCESS

- Users can define genes, pathways, and Gene Ontology terms in their areas of interest to receive automated notifications of curation updates for these entities.
- Extensive web service API provided.
- Author advanced queries: The Structured Advanced Query Form enables intuitive construction of database queries of SQL power using a Web-based interface.

Figure 6: Output from the RouteSearch tool for metabolic path searching. The tool was asked to find paths from aldehyde-D-xylose to ethylene glycol in *E. coli*. The two lowest-cost paths are shown; the second path is truncated. Coloring of the chemical structures indicates conservation of atoms along the pathways. Red arrows indicate reactions added from MetaCyc.

## ABOUT SRI'S BIOINFORMATICS RESEARCH GROUP

SRI International, an independent research institute, is a key player in the field of computational biology, which uses computer science principles and powerful computing capabilities to understand complex biological systems. SRI's Bioinformatics Research Group is a leader in the development of database content and software tools for bioinformatics.

## REFERENCES

- [1] The EcoCyc Database. EcoSal Plus. 2014. <http://www.asmscience.org/content/journal/ecosalplus/10.1128/ecosalplus.ESP-0009-2013ecosalplus.ESP-0009-2013>
- [2] MetaCyc: Nucleic Acids Research 444: D471-80 2016 <http://nar.oxfordjournals.org/content/44/D1/D471.long>
- [3] "Pathway Tools version 19.0: Software for Pathway/Genome Informatics and Systems Biology," *Briefings in Bioinformatics* 2015. <http://arxiv.org/abs/1510.03964>
- [4] "The Pathway Tools Software and Its Role in Anti-Microbial Drug Discovery," *Microbial Genomics and Drug Discovery*, T.J. Dougherty and S.J. Projan eds., Marcel Dekker Inc., New York, 2003

Additional publications: <http://biocyc.org/publications.shtml>

## FOR MORE INFORMATION

For information on subscribing to BioCyc please contact [biocyc@phoenixbioinformatics.org](mailto:biocyc@phoenixbioinformatics.org).

For technical information on BioCyc please contact [biocyc-info@ai.sri.com](mailto:biocyc-info@ai.sri.com).

Visit our website: <http://biocyc.org>

SRI and SRI International are registered trademarks of SRI International. All other trademarks are the property of their respective owners. ©2016. All rights reserved. 04/16

## SRI International

333 Ravenswood Avenue  
 Menlo Park, CA 94025-3493  
 650.859.2000

[www.sri.com](http://www.sri.com)