

Pathway Tools Software and BioCyc.org Genome and Metabolic Pathway Web Portal

SRI International offers innovative tools for modeling and analyzing genomes, metabolic pathways, and regulatory networks to support activities in drug discovery, agriculture, biotechnology, and basic research. These tools accelerate research and support novel analyses, leading to **better science, faster**.

SRI's BioCyc.org web portal contains 20,000 Pathway/Genome Databases (PGDBs) for sequenced genomes. A PGDB contains the entire genome of an organism, as well as its biochemical pathways and (when curated) its regulatory network. Two BioCyc databases, EcoCyc¹ and MetaCyc² are derived from more than three decades of literature-based curation of genome and pathway data. The HumanCyc database provides a unique collection of human metabolic pathway data. The BioCyc.org website is used each year by more than 600,000 unique users worldwide.

The Pathway Tools software is the software underlying BioCyc.³ It is a pathway/genome data management system that enables creation of a private BioCyc portal, and/or creation of BioCyc-like portals containing custom genome data. The software can ingest proprietary genome data, compute a metabolic reconstruction, then make the genome accessible through a large number of search, visualization, and comparative-analysis tools. Data can be accessed using a web browser and through a desktop application. Multiple tools for transcriptomics and metabolomics data analysis are provided. BioCyc and Pathway Tools are separate products but share much functionality; we generally discuss them together but outline differences below.

“For Ginkgo scientists, BioCyc and Pathway Tools are the go-to resources of knowledge and software for exploring ideas, answering questions, and analyzing data.”

— Ginkgo Bioworks

“I'm impressed with the level of detail and care in BioCyc annotations. It is a tremendous resource for pathway analysis in metabolomics.”

— Prof. Art Edison, University of Georgia

BIOCYC FEATURES

- | | |
|--|---|
| <ul style="list-style-type: none"> 20,000 Pathway/Genome Databases 68 databases curated from 146,000 publications Extensive mini-reviews for genes and pathways MetaCyc: Most comprehensive pathway database | <ul style="list-style-type: none"> BioCyc.org Tools: <ul style="list-style-type: none"> Search, visualization, and comparative analysis of genome and pathway information Transcriptomics data analysis Metabolomics data analysis Metabolic route search |
|--|---|

BIOCYC BENEFITS: BETTER SCIENCE, FASTER

Time savings is a major reason users prefer BioCyc over free databases:

- BioCyc curators distill the literature, saving users time searching, reading, synthesizing, and reconciling errors in the literature
- BioCyc integrates a vast array of data, reducing searching of other databases
- Gene functions are more quickly understood in their pathway and regulatory context

BioCyc's accurate, comprehensive data collection and unique analysis tools are a foundation for better science:

- Higher quality publications and grant proposals
- Faster more complete analysis of metabolomics, transcriptomics, and multi-omics data using pathway-based analyses
- Tool for design of optimized pathways for synthetic biology and metabolic engineering

APPLICATIONS OF BIOCYC AND PATHWAY TOOLS

BioCyc and Pathway Tools span genome informatics, pathway informatics, and regulatory informatics. Applications include:

- Enterprise Genome Data Management:** Extensive search tools to speed information finding; multiple visualization tools expedite user uptake of information.
- Computational Inferences:** Predict metabolic pathways, genes coding for missing enzymes in metabolic pathways, protein complexes, and operons, from an annotated genome or metagenome.
- Microbiome Analysis:** Create PGDBs from metagenome-assembled genomes (MAGs); search and compare community members.
- Omics Data Analysis:** Display gene expression, metabolomics, and proteomics data in the Omics Dashboard (Figure 1), on a metabolic map diagram configured for each organism (Figure 2), and on a genome map diagram (Figure 3).
- Drug Discovery:** BioCyc databases facilitate discovery of new drugs through improved pathway-based target selection and validation⁴:
 - Target selection:** Pathway Tools finds drug targets by identifying essential genes using chokepoints and metabolic modeling, identifying enzymes present in multiple pathways, and identifying previously uncharacterized genes filling holes in the metabolic network.
 - Lead generation:** Extensive data on enzyme inhibitors in EcoCyc and MetaCyc
 - Target and lead evaluation:** Improved analysis of omics data
- Metabolic Engineering:**
 - Fast characterization of cellular metabolism for sequenced industrial microorganisms

- RouteSearch Tool: Searches for routes within existing metabolic network; designs novel pathways by combining native reactions with MetaCyc reactions.
- Use quantitative metabolic modeling to guide alternative strain designs
- Comprehensive catalog, through MetaCyc, of known metabolic reactions and metabolic enzymes
- Comparative Genome and Pathway Analyses

THE BIOCYC DATABASE COLLECTION

Each PGDB in BioCyc.org collection the genome and metabolic network of a single organism, and its regulatory network, when curated. PGDBs integrate data imported from external databases, computational inferences, and (for Tier 1 and 2 databases) curated information.

The following are the most highly curated databases in BioCyc:

- **EcoCyc:** Pathway/Genome Database for *Escherichia coli* K-12 MG1655. EcoCyc data have been gathered during two decades of literature-based curation from more than 43,000 articles.¹ EcoCyc provides the equivalent of 4,000 textbook pages of mini-review summaries for 4,172 *E. coli* genes.
- **HumanCyc:** HumanCyc includes literature- based curation of human enzymes and metabolic pathways and is frequently used for metabolomics data analysis.
- **MetaCyc:** Contains 3,100 metabolic pathways and 18,000 reactions from all domains of life. MetaCyc data were gathered from 76,000 publications to provide a comprehensive metabolic encyclopedia; its mini-review summaries encompass the equivalent of 10,600 textbook pages.²

BioCyc PGDBs are organized into three tiers:

- **Tier 1 databases** EcoCyc, MetaCyc, HumanCyc, YeastCyc, and AraCyc (*Arabidopsis*) have received extensive manual curation.
- **Tier 2 databases** are computationally generated from RefSeq entries, followed by up to one person-year of subsequent curation. Tier 2 databases include *Bacillus subtilis*,

Saccharomyces cerevisiae, *Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Clostridioides difficile*, *Streptococcus pneumoniae*, *Helicobacter pylori*, and *Vibrio cholerae*.

- **Tier 3 databases** are computationally generated from RefSeq entries with no subsequent curation.

All PGDBs include the genome, predicted metabolic pathways, predicted pathway hole fillers – genes coding for missing (unannotated) enzymes in metabolic pathways – and, for bacteria, predicted operons.

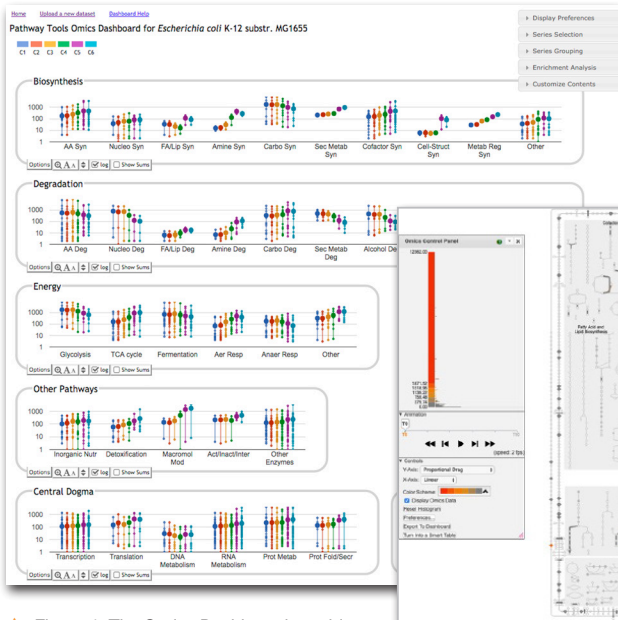
Additional types of data integrated in BioCyc include:

- Protein subcellular locations, enzyme kinetics data, protein features, Gene Ontology terms, predicted Pfam domains
- Regulatory information including promoters, operons, transcription-factor binding sites, regulatory RNAs
- Gene essentiality data
- Reaction atom mappings
- Gibbs free energies of formation for metabolites
- Growth media
- Ortholog relationships to other BioCyc genomes
- Database links (e.g., to UniProt and RefSeq)

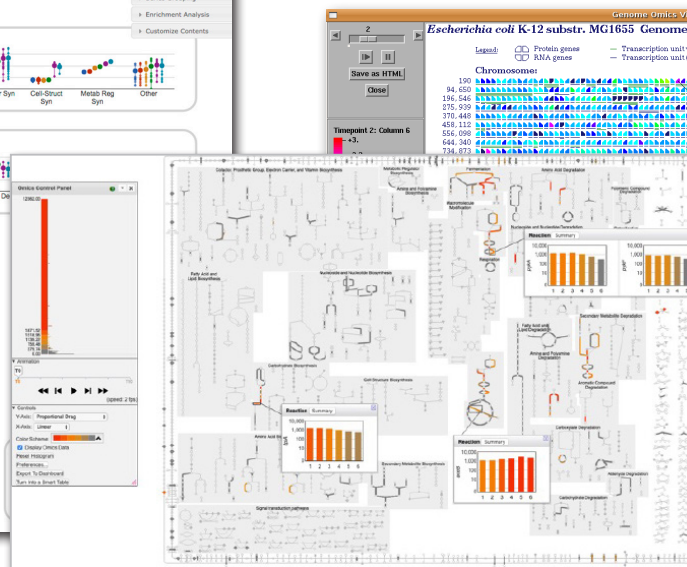
GENOME INFORMATICS TOOLS

- Search for genomes by name, taxonomy, phenotypic properties
- Gene information page
 - Retrieve amino-acid sequence and nucleotide sequence of arbitrary genome region
 - Search genes by name, accession number, sequence length, replicon position, GO terms, and protein properties (pI, MW, protein features, subcellular location, ligand)
- The Genome Browser (Figure 4) was re-implemented in 2023 for greater speed and functionality. It depicts genomic regions at user-selected resolution with semantic zooming. Visible features include pseudogenes, promoters, transcription-factor binding sites, repeats, terminators, and nucleotide sequence. Genome posters can be generated.
- Transcription-unit information page

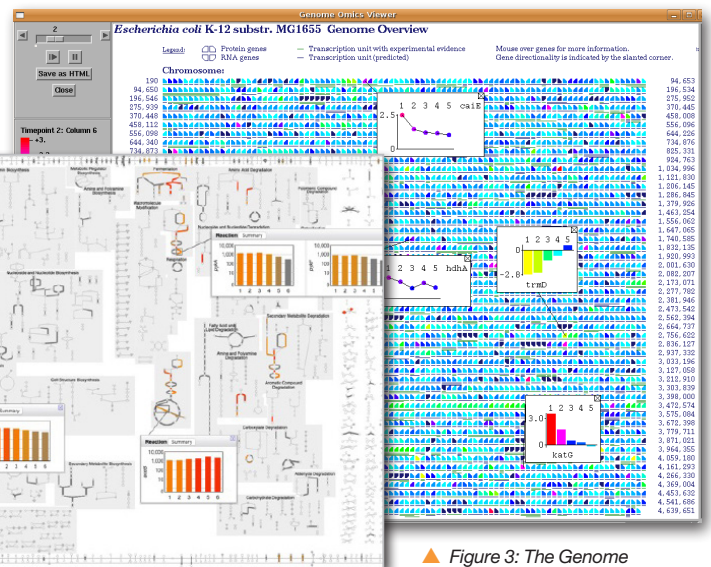
BLAST search, sequence-pattern searching, map SNPs to genes and show effects on translation, sequence alignment.



▲ Figure 1: The Omics Dashboard provides a visual read-out of the cellular state within an omics dataset.



▲ Figure 2: 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 3: The Genome Omics Viewer depicts an entire chromosome in one window. Genes are colored according to a gene expression dataset.

PATHWAY AND REGULATORY INFORMATICS TOOLS

- Pathway information page
 - Each pathway shows a detailed mini-review from MetaCyc
 - Search pathways by name, substrates, and length
- Reaction information page includes atom mappings
- Metabolite information page
 - Search metabolites by name, accession numbers, substructure, mass, monoisotopic mass, elemental composition
- Customize pathway diagrams for publication or assemble groups of pathways into a pathway collage.
- Cellular Overview diagrams (Figure 2) are organism-specific depictions of metabolic and transporter networks that are zoomable and searchable.
- Route Search tool finds minimum-cost paths between metabolites in the metabolic network. Route Search paths maximize the number of atoms conserved from source metabolite to target metabolite by using an extensive library of reaction atom mappings. Route Search can find paths across multiple organisms from a microbiome.
- The regulatory Overview presents the genetic regulatory network stored in a PGDB.

ANALYSIS TOOLS FOR GENE EXPRESSION AND METABOLOMICS DATA

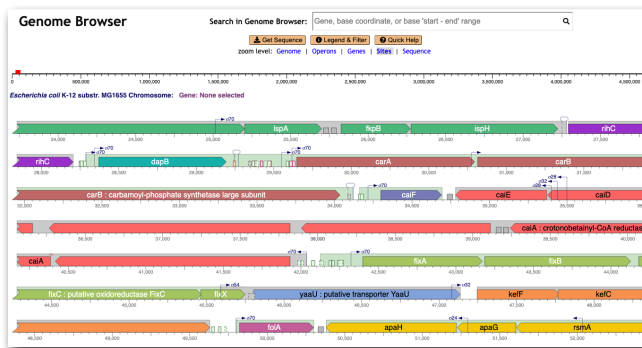
- SmartTables store lists of genes, metabolites, pathways, and more. Browse database attributes, share with colleagues, transform to pathway lists, perform enrichment analysis.
- The Omics Dashboard presents a visual read-out of the expression status of all cellular systems to facilitate a rapid top-down user survey of cellular responses (Figure 1)
- Paint omics data onto individual pathways, pathway collage, and full metabolic networks (Figure 2) (example animation at <http://biocyc.org/ov-expr.shtml>).

COMPARATIVE GENOMICS TOOLS

- Comparative genome browser (Figure 5) 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, and pathways across organisms

ADVANCED DATABASE ACCESS

- Web services API provided.
- Python, Perl, Java, and Lisp APIs.



▲ Figure 4. Genome browser depiction of a region of the *E. coli* chromosome. Gene colors indicate operon organization. Promoters, transcription factor binding sites, and terminators are depicted when known.

- Author advanced queries of SQL power using the intuitive Structured Advanced Query Form.
- Interoperability: Export PGDBs to the BioPAX, SBML, GFF, and Genbank formats.

PATHWAY TOOLS SOFTWARE

The Pathway Tools software powers the BioCyc.org web portal. Bring Pathway Tools in-house to create a private BioCyc portal containing selected BioCyc PGDBs, and to create PGDBs for proprietary genomes.

METABOLIC MODELING IN PATHWAY TOOLS

The MetaFlux module generates a flux balance analysis (FBA) model automatically from a PGDB, enabling quantitative modeling of steady-state metabolic fluxes. Combine models for multiple organisms to model organism communities.

Simulate spatial interactions by situating organisms within a spatial grid. Dynamic FBA enables temporal simulations.

By combining pathway databases with FBA, MetaFlux achieves close coupling of the FBA model to genome and metabolite data; and high accessibility of the FBA model via the query and visualization features of Pathway Tools. MetaFlux speeds comprehension of simulation results by painting computed fluxes on metabolic-map diagrams and on individual pathways, and MetaFlux will plot metabolite concentrations and organism biomass changes.

Development of FBA models is accelerated by a multiple gap-filling tool that postulates additional reactions to add to an FBA model to complete it, and that identifies what subset of biomass components can be produced by the current model. Modeling of gene knock-outs is supported.

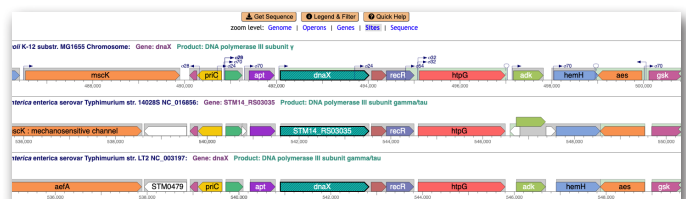
PATHWAY TOOLS BENEFITS

Bring BioCyc into a secure proprietary environment
Extract more information from proprietary genomes via

- Computational inferences
- Search and visualization tools
- Comparative tools
- Omics data analysis tools

PATHWAY TOOLS FEATURES

- Enterprise pathway/genome data management
- Desktop and web operation
- Computational prediction of metabolic pathways, operons, protein complexes
- Incorporate selected BioCyc databases
- Quantitative metabolic modeling
- Includes BioCyc tools



▲ Figure 5. Comparative genome browser showing alignments with respect to the *dnaX* gene in one *E. coli* genome, and two *Salmonella* genomes. Colors indicate orthologous gene groups.

TECHNICAL SPECIFICATIONS/CONFIGURATIONS

1. Pathway/Genome Navigator Bundled with BioCyc Databases

All configurations provide query, visualization, and analysis of existing BioCyc databases. The same binary application can run as both a desktop application and as a Web server within an organization's intranet.

Configurations available:

1) Pathway/Genome Navigator plus as many as 500 user-selected BioCyc databases

- Platforms supported: Linux, Windows (Web mode not available), Macintosh
- Hardware: 2.4 GHz processor, 16 GB RAM, 5 GB disk

2) Pathway/Genome Navigator plus all BioCyc databases

- Platforms supported: Linux
- Hardware: 2.4 GHz processor, 128 GB RAM, 3 TB disk

2. Add Editors, PathoLogic, MetaFlux

- To edit existing BioCyc PGDBs, add the Pathway/Genome Editors alone to configuration (1).
- To create and edit new Pathway/Genome Databases, add PathoLogic and Pathway/Genome Editors to configuration (1).
- PathoLogic requires the EcoCyc and MetaCyc PGDBs.
- To run quantitative metabolic flux models, add MetaFlux to configuration (1).

COMPREHENSIVE LIST OF FEATURES IN THE PATHWAY TOOLS SOFTWARE SUITE

Genome Informatics

- Gene/protein/RNA searches, gene information page
- Genome browser
- BLAST search, sequence pattern search
- Multiple sequence alignment
- Sequence retrieval

Pathway Informatics

- Reaction, metabolite, pathway searches
- Reaction, metabolite, pathway pages
- Single and multi-pathway diagrams
- Zoomable metabolic network diagrams
- Pathway inference from annotated genome
- Quantitative metabolic modeling via flux- balance analysis: FBA, dFBA, FVA, gap filling, blocked metabolites
- Metabolic route search for metabolic engineering
- Metabolic route search for microbial communities
- Metabolic network explorer

Regulatory Informatics

- Operon inference from genome
- Capture/visualize promoters, transcription factor binding sites
- Visualize regulon of transcription factor
- Regulatory network viewer

Transcriptomics Data Analysis

- Enrichment analysis
- Paint transcriptomics data onto single pathways, zoomable metabolic network

- Paint transcriptomics data onto regulatory network
- Paint transcriptomics data onto genome diagram
- Omics Dashboard
- Sort pathways by pathway activation score
- Multi-omics Explainer

Metabolomics Data Analysis

- Enrichment analysis
- Paint metabolomics data onto single pathways, zoomable metabolic network
- Omics Dashboard
- Sort pathways by pathway activation score
- Pathway covering analysis

Microbiome Informatics

- Calculate pathway abundances across metagenome samples
- Create PGDBs for an organism community
- Search across multiple PGDBs, compare PGDBs
- Analysis of meta-omics data

Advanced Data Manipulation

- SmartTables
- Interactive editing tools
- Advanced Query Interface

APIs and Data Import/Export

- APIs: Web services, Python, Java, Perl, R, Lisp
- Import/Export: SBML, GenBank, GFF, BioPAX, tab-delimited files

ABOUT SRI'S BIOINFORMATICS RESEARCH GROUP

SRI International, an independent research institute, is a key player in the field of computational biology. SRI's Bioinformatics Research Group, which produces BioCyc and Pathway Tools, is a leader in the development of database content and software tools for bioinformatics.

REFERENCES

[1] The EcoCyc Database. EcoSal Plus. 2018. <http://www.asmscience.org/content/journal/ecosalplus/10.1128/ecosalplus.ESP-0006-2018>

[2] MetaCyc. Nucleic Acids Research 2020. doi: 10.1093/nar/gkz862

[3] Pathway Tools version 24.0: Integrated Software for Pathway/Genome Informatics and Systems Biology, 2020. <http://arxiv.org/abs/1510.03964v4>

[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

Downloadable Pathway Tools Software

- Freely available to academics and government laboratories for research purposes; see <http://biocyc.org/download.shtml>
- For commercial use click here <https://share.hsforms.com/13fwJvezjSDyUXI8BRPpkQ3ipje> or contact ptools-commercials@ai.sri.com.

BioCyc.org Website Subscriptions

- Subscriptions provide access to more than 20,000 organism databases and extensive online software tools.
- The EcoCyc database is free to all users.
- Subscriptions are available for institutions, labs, and individuals.
- Contact biocyc-sales@sri.com.

SRI International

333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000
www.sri.com