SRI International

Pathway Tools | BioCyc Database Collection

BioCyc Database Collection and Pathway Tools Software

SRI International offers innovative tools for modeling and analyzing genomes, metabolic pathways, and regulatory networks to support activities in drug discovery, agriculture, and biotechnology. These tools accelerate research and lead to a greater understanding of biological systems. SRI's unique Omics Viewers support visualization and analysis of large omics datasets on genome-scale cellular network diagrams. BioCyc and Pathway Tools are freely available for academic research.

The BioCyc database collection from SRI International is a set of 3,000 Pathway/Genome Databases (PGDBs) for many sequenced genomes. PGDBs describe the entire genome of an organism, as well as its biochemical pathways and (when curated) its regulatory network. New expanded releases occur twice per year. 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 unique collection of human metabolic pathway data.

The downloadable Pathway Tools software (licensed to date to more than 4,500 groups) and BioCyc databases provide many capabilities not present on the BioCyc.org Web site.

> *"Pathway Tools is a valuable component of our systems biology toolbox. The software allows me to integrate and visualize various types of omics data within the context of the metabolic network for the bacterial and fungal strains we employ. Further, the software provides a huge benefit in allowing us to custom-tailor Pathway/Genome Databases to our specific strain(s)."*
>
> Randy Berka
> Director, Novozymes, Inc.

APPLICATIONS OF BIOCYC AND PATHWAY TOOLS

Pathway/Genome Databases are platforms for knowledge management and data analysis for organisms of critical importance for R&D activities.

- **Genome and Metagenome Analysis:** Predict metabolic pathways, genes coding for missing enzymes in metabolic pathways, and operons, from an annotated genome.
- **Metabolic flux analysis:** Fast generation of flux models using flux-balance analysis.

KEY CAPABILITIES OF PATHWAY TOOLS AND BIOCYC

- **Develop a comprehensive, genome database for mission-critical organisms**
- **Computational prediction of metabolic pathways, missing enzymes, operons**
- **Omics data analysis using genome-scale visualizations of the metabolic network, regulatory network, and genome**
- **Steady-state metabolic modeling using flux-balance analysis**

RECENT ENHANCEMENTS

- Metabolic route search and pathway design tool
- SmartTables speed interactive data analysis
- Retrieve gene expression data from GEO and PortEco
- Support for Phenotype Microarray data
- Atom mappings present for most metabolic reactions
- Sequence pattern searches and multiple alignments
- Support for glycans and glycan pathways
- New web services

PATHWAY TOOLS COMPONENTS

- PathoLogic: Computational inference tools for predicting pathways, pathway hole fillers, and operons
- MetaFlux: Flux Balance Analysis
- RouteSearch: Pathway design and metabolic route finding
- Pathway/Genome Navigator: Supports querying, visualization, and analysis of PGDBs
- Pathway/Genome Editors: Interactive editing of PGDBs

- **Omics Data Analysis:** Paint combinations of gene expression, metabolomics, and proteomics data onto a metabolic map diagram uniquely configured for each organism (Figure 1), onto a cellular regulatory network (Figure 2), and onto a genome map diagram (Figure 3). Genome Browser tracks speed analysis of ChIP-chip data and other positional data.
- **Encyclopedic Reference:** Query and display relationships among genome and pathway data.
- **Drug Discovery:** BioCyc databases facilitate discovery of new drugs through improved pathway-based target selection and validation[4]:
  - **Target selection:** Pathway Tools algorithms for finding drug targets include identifying essential genes using 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

BioCyc and Pathway Tools are freely available to academic groups for research purposes.

- Metabolic Engineering:
  - Fast characterization of cellular metabolism for industrial microorganisms
  - **RouteSearch tool:** designs pathways by combining native reactions with MetaCyc reactions
  - Use quantitative metabolic modeling to guide alternative strain designs
  - Tracking of alternative strain designs
  - Comprehensive catalog, through MetaCyc, of known metabolic reactions and metabolic enzymes
- **Comparative Analyses:** Comparative genome analyses and comparative pathway analysis can be performed.

THE BIOCYC DATABASE COLLECTION

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

- **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 26,000 articles[1]. EcoCyc provides mini-review summaries for 3,800 *E. coli* genes, and descriptions of the metabolic and transcriptional regulatory networks of *E. coli*.
- **HumanCyc:** Developed from a computational pathway analysis of the human genome, literature-based curation of HumanCyc resumed in 2009 and is ongoing.
- **MetaCyc:** Contains 2,100 metabolic pathways and 11,700 reactions from 2,500 organisms in all domains of life. MetaCyc data and commentary were gathered from 40,000 publications to provide a comprehensive metabolic encyclopedia[2].

The databases are organized into three tiers.

- **Tier 1 Pathway/Genome Databases** have received extensive manual curation, and include EcoCyc, MetaCyc, and HumanCyc.
- **Tier 2 Pathway/Genome Databases** were computationally generated with less than one person-year of subsequent curation, and include databases for *Bacillus subtilis, Agrobacterium tumefaciens, Bacillus anthracis, Francisella tularensis, Helicobacter pylori, Mycobacterium tuberculosis, Vibrio cholerae,* and others.
- **Tier 3 Pathway/Genome Databases** were computationally generated with no subsequent curation.

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

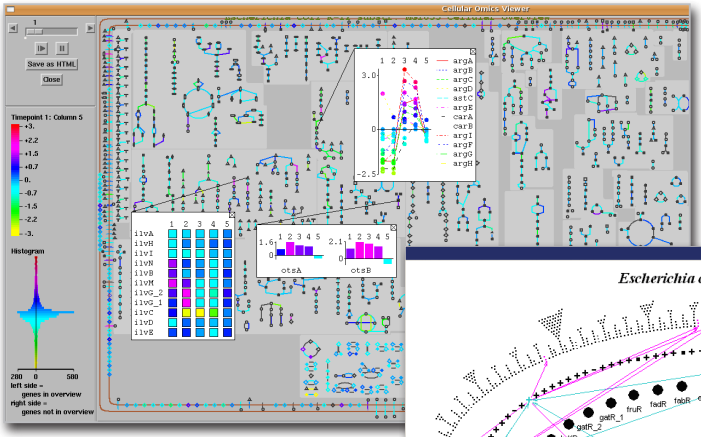
PATHWAY TOOLS SOFTWARE

Pathway Tools[4] provides a powerful and comprehensive set of features for querying, visualization, analysis, and curation of the BioCyc database collection. Pathway Tools combines representation and inference techniques from artificial intelligence to extract additional information from genomes, and encode that information within a sophisticated ontology to enable computational manipulation.

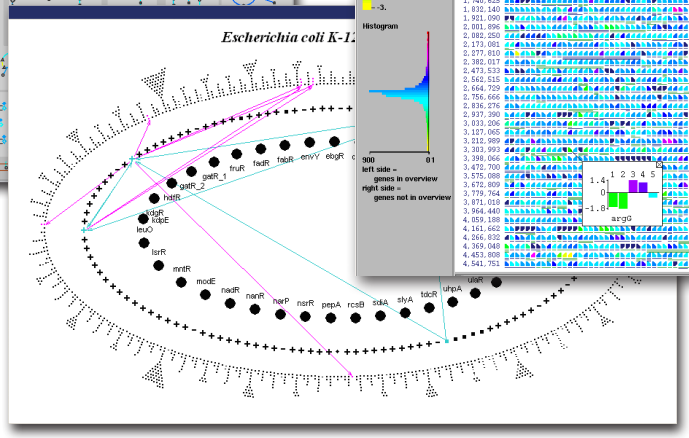
Pathway Tools can operate on the BioCyc collection of PGDBs available through SRI, and on locally created PGDBs, such as for proprietary genomes. It can also operate on more than 600 PGDBs created by third parties, including those listed in SRI's *Registry of Pathway/Genome Databases* at *http://biocyc.org/registry.html.*

Querying, Visualization, and Analysis Of Pathway and Genome Databases

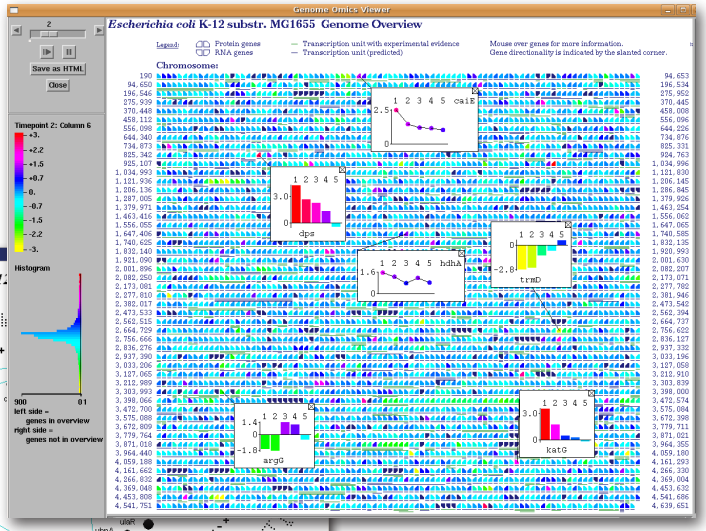
The Pathway/Genome Navigator runs in both a Web mode that allows publishing of a PGDB on the Internet or an intranet, and as a desktop application. Features include



▲ *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.*

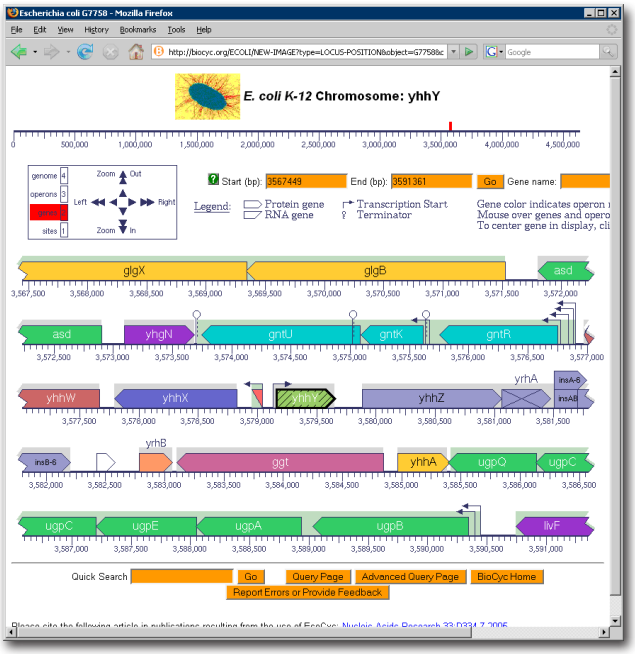


▲ *Figure 2: The Regulatory Overview depicts the transcriptional regulatory network in a PGDB. Here, the E. coli regulatory network is highlighted to show genes that regulate the gntR gene (blue lines), and the genes that are regulated by gntR (purple lines). The inner two rings are populated by transcription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



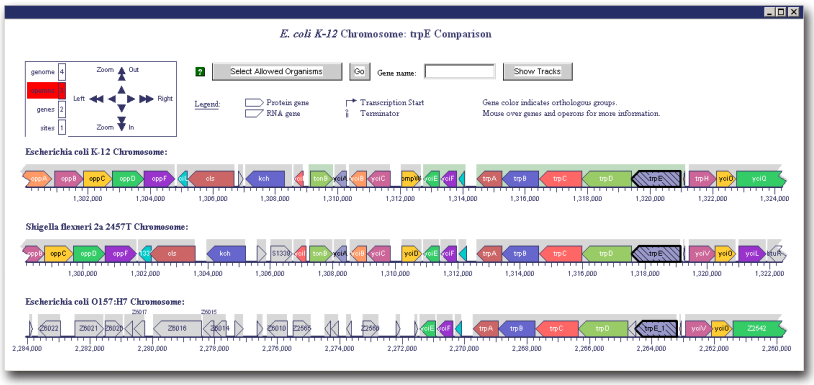
▲ *Figure 3: The Genome Omics Viewer depicts an entire chromosome in one window. Genes are colored according to a gene expression dataset.*

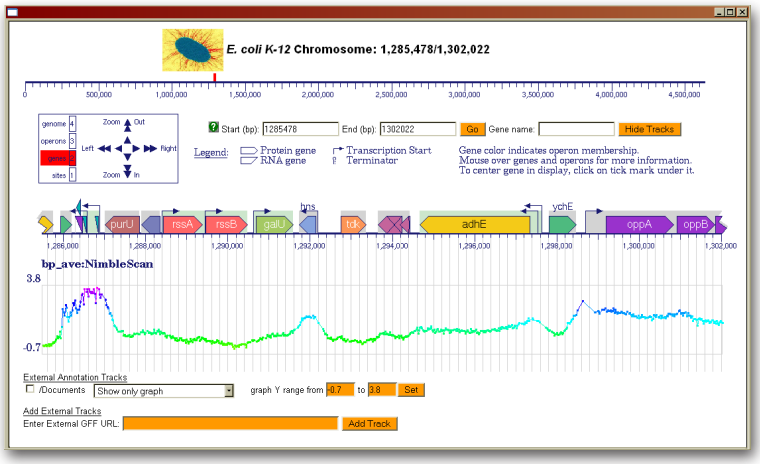
- **Visualizations tailored to each of the following object types:**
  - Biochemical pathways and reactions
  - Chemical compounds
  - Genes, proteins, and RNAs
  - Operons and regulatory interactions
- **Genome Browser** (Figure 4) depicts genomic regions at user-selected resolution with semantic zooming that reveals new features at higher resolutions. **Genome Omics Viewer** (Figure 3) paints omics data onto a low-resolution depiction of an entire genome. **Tracks** facility (Figure 5) allows user datasets to be plotted against the genome.
- **Regulatory Overview** (Figure 2) presents the genetic regulatory network stored in a PGDB. **Regulatory Omics Viewer** paints omics datasets onto the regulatory network to enable comparisons of expression measurements with regulatory mechanisms.
- **Cellular Overview** diagram (Figure 1) is a full-screen depiction of the metabolic and transporter networks of an organism; it provides many tools for navigating the networks.
- **Cellular Omics Viewer** (Figure 1) enables the user to paint omics datasets onto the Cellular Overview diagram, together or individually. Scientists can interpret gene expression, proteomics, and metabolomics datasets in a pathway context, including animation of time-course or comparative datasets (example animation at *http://biocyc.org/ ov-expr.shtml*).
- **All omics viewers** (Figure 1-3) support animation of time course or comparative omics datasets, and graphing of individual omics data values.
- **Over-representation analysis** of gene sets and metabolite sets for gene ontology categories, regulons, and pathways.
- **Comparative genomics operations** (Figure 4) support a range of genome and metabolic comparisons, from visualization of chromosomal regions around orthologous genes to comparisons of whole metabolic networks.
- **Metabolic map posters and genome posters** are automatically generated from a PGDB.
- **Anti-microbial drug targets** are predicted by a tool that computes metabolic choke points.
- **RouteSearch tool** finds minimum-cost paths between metabolites in the metabolic network, and designs pathways to new metabolites by importing reactions from MetaCyc.
- **SmartTables** enable users to store and analyze groups of genes, metabolites, pathways, etc. Data are presented as spreadsheets. Analyses include transformations (e.g., transform a pathway table to a table of all genes in the pathways, or a metabolite table to a table of all pathways containing those metabolites), over-representation analysis, and manipulation of sequence regions.
- **Dead-end metabolite** identification algorithm provided.
- **Author advanced queries** Interact with the Structured Advanced Query Form, which supports intuitive construction of database queries of SQL power using a Web-based interface.
- **Interoperability** is supported by exporting PGDBs to the BioPAX, SBML, and Genbank formats. PGDBs can be queried through Perl, Java, and Lisp APIs via web services, and PGDBs, and be imported into SRI's open source BioWarehouse system (Oracle or MySQL) for integration and cross-querying with multiple bioinformatics databases.



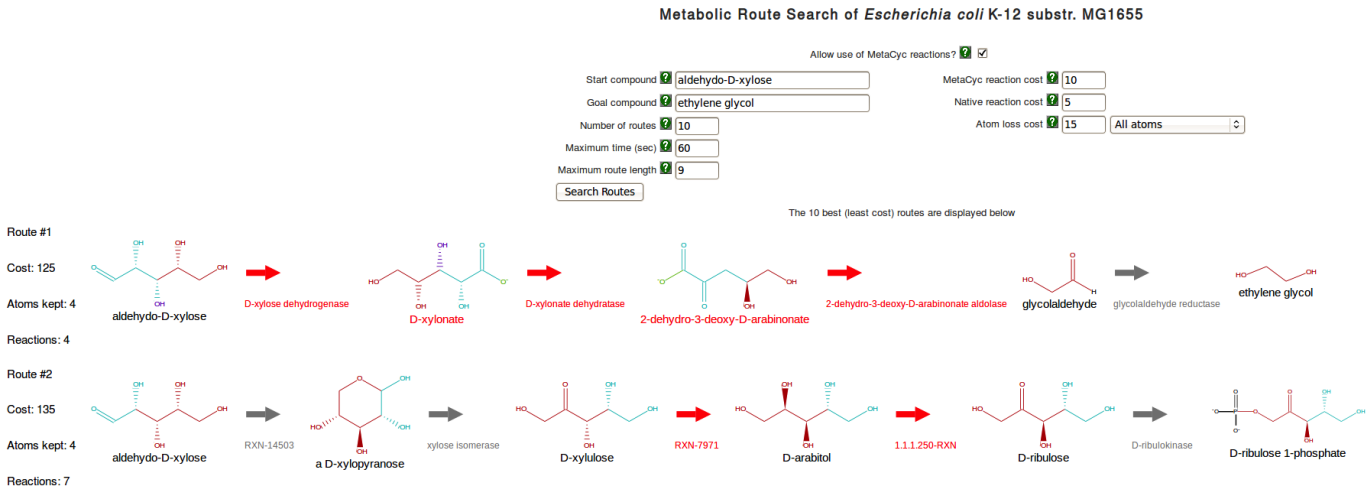
▶ *Figure 5: Genome browser with tracks display enabled. The single track shown here was generated from a data file containing ChIP-chip data for RNA polymerase binding. This facility allows the user to compare the frequency of protein binding from ChIP-chip experiments against curated promoters within a PGDB.*

◀▼ *Figure 4. Left: Genome browser depiction of a region of the E. coli chromosome. Gene colors indicate operon organization. Promoters and terminators are depicted when known. Pseudogenes are marked with X's. Below: Comparative genome browser showing alignments with respect to the trpE gene of two E. coli genomes and the Shigella flexneri genome. Colors indicate orthologs.*





Generate an organism-specific metabolic wall chart from a PGDB.



▲ *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.*

Flux-Balance Analysisb (FBA) in Pathway Tools

The MetaFlux FBA Module generates an FBA model automatically from a PGDB, enabling quantitative modeling of steady-state metabolic fluxes. By coupling pathway databases with FBA, we achieve closer coupling of the FBA model to genome information; more accessibility of the FBA model via the query and visualization features of Pathway Tools; and accelerate comprehension of FBA results by painting computed fluxes on the cellular overview diagram and on individual pathways.

Development of functional FBA models is greatly 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.

TECHNICAL SPECIFICATIONS/CONFIGURATIONS

1. Pathway/Genome Navigator Bundled with BioCyc Databases

These 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 your organization's intranet.

Configurations available:

1) **Pathway/Genome Navigator plus as many as 20 user-selected BioCyc databases**
   - Platforms supported: Linux/x86, Windows/x86, Macintosh
   - Hardware: 3 GHz processor, 2 GB RAM, 5 GB disk
2) **Pathway/Genome Navigator plus all BioCyc databases**
   - Platforms supported: Linux/x86 64bit
   - Hardware: 3 GHz processor, 64 GB RAM, 500 GB disk

2. Add Pathway/Genome Editors and PathoLogic

- 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. In the PathoLogic configuration, the EcoCyc and MetaCyc PGDBs are also required.
- Use of Pathway/Genome Editors optionally involves the MySQL relational database system

ABOUT SRI'S BIOINFORMATICS RESEARCH GROUP

SRI International, an independent research institute, is a key player in the emerging 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] EcoCyc: Nucleic Acids Research 41:D605-12 2013 *http://nar.oxfordjournals.org/content/41/D1/D605.long*
[2] MetaCyc: *Nucleic Acids Research* 42:D459-71 2014 *http://nar.oxfordjournals.org/content/42/D1/D459.long*
[3] "Pathway Tools version 13.0: Integrated Software for Pathway/Genome Informatics and Systems Biology," *Briefings in Bioinformatics 2010 11:40-79. http://www.ai.sri.com/pkarp/misc/ptools09.html*
[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

Academic and research organizations: BioCyc and Pathway Tools are available free of charge under license. See *http://biocyc.org/downloads.shtml* for more information.

Commercial organizations: Contact Doug Bercow, Director, Business Development, SRI International (www.sri.com) at (650) 859-5187 or *douglas.bercow@sri.com*.

Visit our website: *http://biocyc.org*

SRI International

333 Ravenswood Avenue
Menlo Park, CA 94025-3493
650.859.2000

www.sri.com