

The Pathway Tools Software and BioCyc Database Collection

Peter D. Karp, Ph.D.

Bioinformatics Research Group

SRI International

pkarp@ai.sri.com

<http://www.ai.sri.com/pkarp/talks/>

BioCyc.org

EcoCyc.org, MetaCyc.org, HumanCyc.org

Use Cases for Pathway Tools and BioCyc

- **Development of organism-specific DBs (model-organism DBs) that span many biological datatypes**
- **Web publishing of those DBs with a powerful set of query and visualization tools**
- **Computational inferences of metabolic pathways, pathway hole fillers, operons, transport reactions**
- **Visual tools for analysis of omics data**
- **Tools for analysis of biological networks**
- **Comparative analysis tools**
- **Metabolic engineering**
- **BioCyc is a Web portal for genome and pathway information**

BioCyc Collection of 673 Pathway/Genome Databases

● Pathway/Genome Database (PGDB) – combines information about

- Pathways, reactions, substrates
- Enzymes, transporters
- Genes, replicons
- Transcription factors/sites, promoters, operons

● Tier 1: Literature-Derived PGDBs

- MetaCyc
- EcoCyc -- *Escherichia coli* K-12

● Tier 2: Computationally-derived DBs, Some Curation -- 28 PGDBs

- HumanCyc
- Mycobacterium tuberculosis

● Tier 3: Computationally-derived DBs, No Curation -- 643 DBs

BioCyc Home - Mozilla Firefox

File Edit View History Bookmarks Tools Help

BioCyc Database Collection

Pathway Tools Workshop August 19-28, 2009 in Menlo Park, CA

Logged in as pkarp@ai.sri.com | Logout | Help | My preferences

Search Database *Escherichia coli* K-12 substr. MG1655 change

Home Search Tools Help

News

BioCyc version 13.1 contains 507 genomes. Read more.

Information

Introduction to BioCyc
Guide to BioCyc
Webinars
507 Databases
Guided Tour
Pathway Tools Software
Publications
Linking to BioCyc
External Links

Services

Join BioCyc Mailing List
Metabolic Posters: **NEW**
Genome Posters: **NEW**
Software/Database Downloads
Registry

ABOUT BIOCYC

BioCyc is a collection of 507 Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism.

To learn more about BioCyc, read the Introduction to BioCyc or watch our free online instructional videos.

BIOCYC TOOLS

The BioCyc Web site contains many tools for navigating and analyzing these databases, and for analyzing omics data, including the following.

- Genome browser
- Display of individual metabolic pathways, and of full metabolic maps
- Visual analysis of user-supplied omics datasets by painting onto metabolic map, regulatory map, and genome map
- Comparative analysis tools

The downloadable version of BioCyc that includes the Pathway Tools software provides more speed and power than the BioCyc Web site [more]. Multiple database configurations are available for installation with the software including multiple *E. coli* and *Shigella* genomes, multiple *Bacillus* genomes, multiple *Mycobacterium* genomes, and multiple mammalian genomes.

BIOCYC PATHWAY/GENOME DATABASES

The BioCyc databases are divided into three tiers, based on their quality.

Tier 1 databases have received person-decades of literature-based curation, and are the most accurate. Tier 2 and Tier 3 databases contain computationally predicted metabolic pathways, predictions as to which genes code for missing enzymes in metabolic pathways, and predicted operons.

PGDBs for many other organisms are available outside the BioCyc collection, created by other users of Pathway Tools. Some of these PGDBs are highly curated, and exist for important model organisms including *Mouse*, *Arabidopsis*, and *Yeast*. For more information on accessing these PGDBs, click here.

BioCyc Tier 1: Intensively Curated Databases

Pathway Tools Software

- **PathoLogic**

- Predicts operons, metabolic network, pathway hole fillers, from genome
- Computational creation of new Pathway/Genome Databases

- **Pathway/Genome Editors**

- Distributed curation of PGDBs
- Distributed object database system, interactive editing tools

- **Pathway/Genome Navigator**

- WWW publishing of PGDBs
- Querying, visualization of pathways, chromosomes, operons
- Analysis operations
 - ◆ Pathway visualization of gene-expression data
 - ◆ Global comparisons of metabolic networks

Briefings in Bioinformatics 11:40-79 2010

Obtaining a PGDB for Organism of Interest

- Find existing curated PGDB
- Find existing PGDB in BioCyc
- Create your own
- Curated pathway DBs now exist for most biomedical model organisms

Pathway Tools Software: PGDBs Created Outside SRI

- 2,100+ licensees: 180 groups applying software to 1,600 organisms
- **Saccharomyces cerevisiae**, SGD project, Stanford University
 - 135 pathways / 565 publications
- **Candida albicans**, CGD project, Stanford University
- dictyBase, Northwestern University

- **Mouse**, MGD, Jackson Laboratory
- **Drosophila**, FlyBase, Harvard University
- Under development:
 - *C. elegans*, WormBase

- **Arabidopsis thaliana**, TAIR, Carnegie Institution of Washington
 - 288 pathways / 2282 publications
- **PlantCyc**, Carnegie Institution of Washington
- **Six Solanaceae species**, Cornell University
- **GrameneDB**, Cold Spring Harbor Laboratory
- **Medicago truncatula**, Samuel Roberts Noble Foundation

MetaCyc: Metabolic Encyclopedia

- Describe a representative sample of every experimentally determined metabolic pathway
- Describe properties of metabolic enzymes
- Literature-based DB with extensive references and commentary
- MetaCyc now assigns more than twice as many reactions to pathways as does KEGG

Nucleic Acids Research 2010

MetaCyc Data -- Version 14.0

Pathways	1,471
Reactions	8,409
Enzymes	6,198
Small Molecules	8,572
Organisms	1,861
Citations	22,459

Taxonomic Distribution of MetaCyc Pathways – version 13.1

Bacteria	883
Green Plants	607
Fungi	199
Mammals	159
Archaea	112

Pathway Tools Survey Publication

- **Karp et al, *Briefings in Bioinformatics* 2010 11:40-79.**

Signaling Pathway Editor

- **Signaling pathways use different visual conventions than metabolic pathways**
- **Look and feel based of our tool based on CellDesigner, SBGN**
- **Manual layout**
 - Can't yet be included in Cellular Overview Diagram

Homo sapiens Pathway: MAP kinase cascade

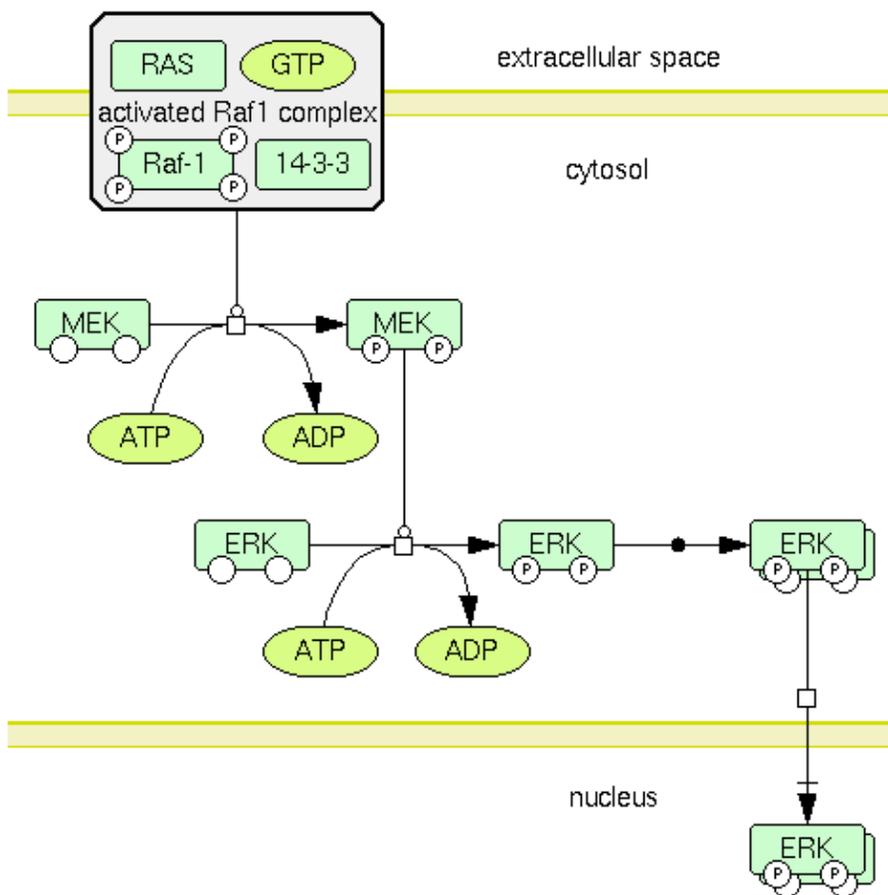


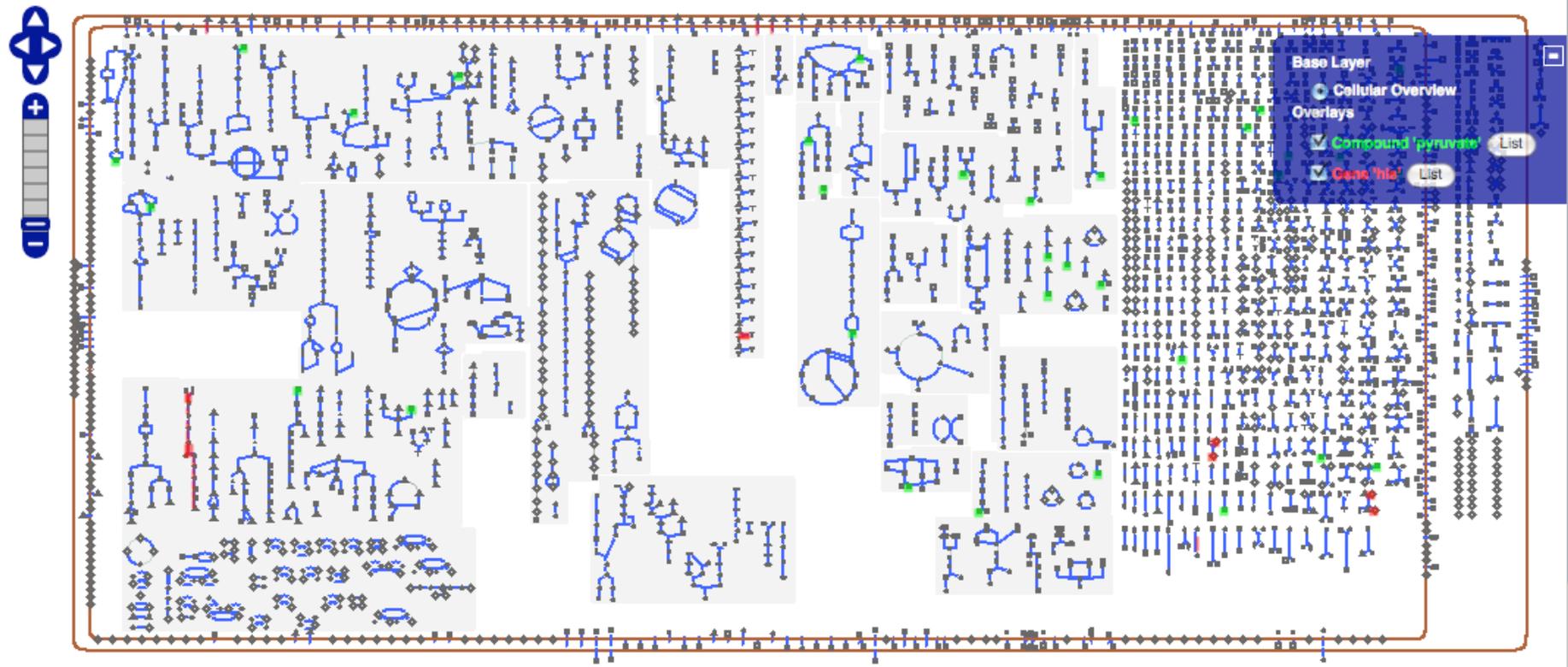
Diagram Key: ?

Improved Web Overviews

- **Implemented using OpenLayers**
- **Zoomable, draggable, searchable, paintable**
- **Cellular Overview**
 - Highlight compounds, reactions, enzymes, genes by name, substring, with autocomplete
 - Highlight genes from file
 - Superimpose omics data
- **Regulatory Overview**
 - Draw connections between a gene and its regulators, regulatees
 - Show full diagram or only highlighted genes

Cellular Overview

Cellular Overview of *Escherichia coli* K-12 substr. MG1655 (Pan left/right/up/down by holding the left button mouse)



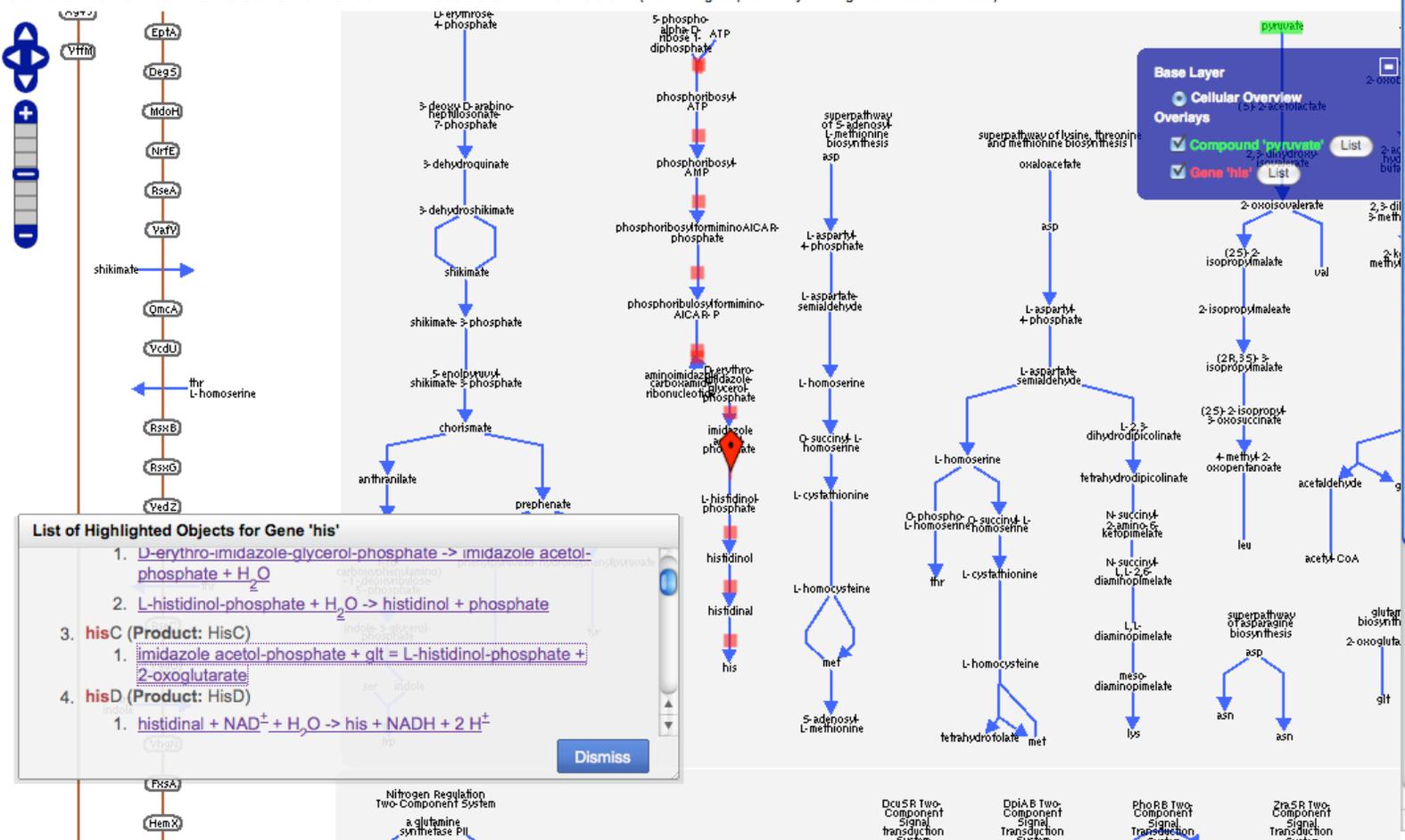
Cellular Overview, zoomed-in view


Pathway Tools Workshop
October 25 - 29, 2010
Menlo Park, CA
LOGIN | Why Login? | Create New Account
Quick Search | Gene Search
Search Database *Escherichia coli* K-12 substr. MG1655 [change](#)

[Home](#) | [Search](#) | [Tools](#) | [Help](#) | [Cellular Overview](#)

Cellular Overview of *Escherichia coli* K-12 substr. MG1655

(Pan left/right/up/down by holding the left button mouse)



List of Highlighted Objects for Gene 'his'

- [D-erythro-imidazole-glycerol-phosphate -> imidazole acetol-phosphate + H₂O](#)
- [L-histidinol-phosphate + H₂O -> histidinol + phosphate](#)
- hisC (Product: HisC)**
 - [imidazole acetol-phosphate + glt = L-histidinol-phosphate + 2-oxoglutarate](#)
- hisD (Product: HisD)**
 - [histidinal + NAD⁺ + H₂O -> his + NADH + 2 H⁺](#)

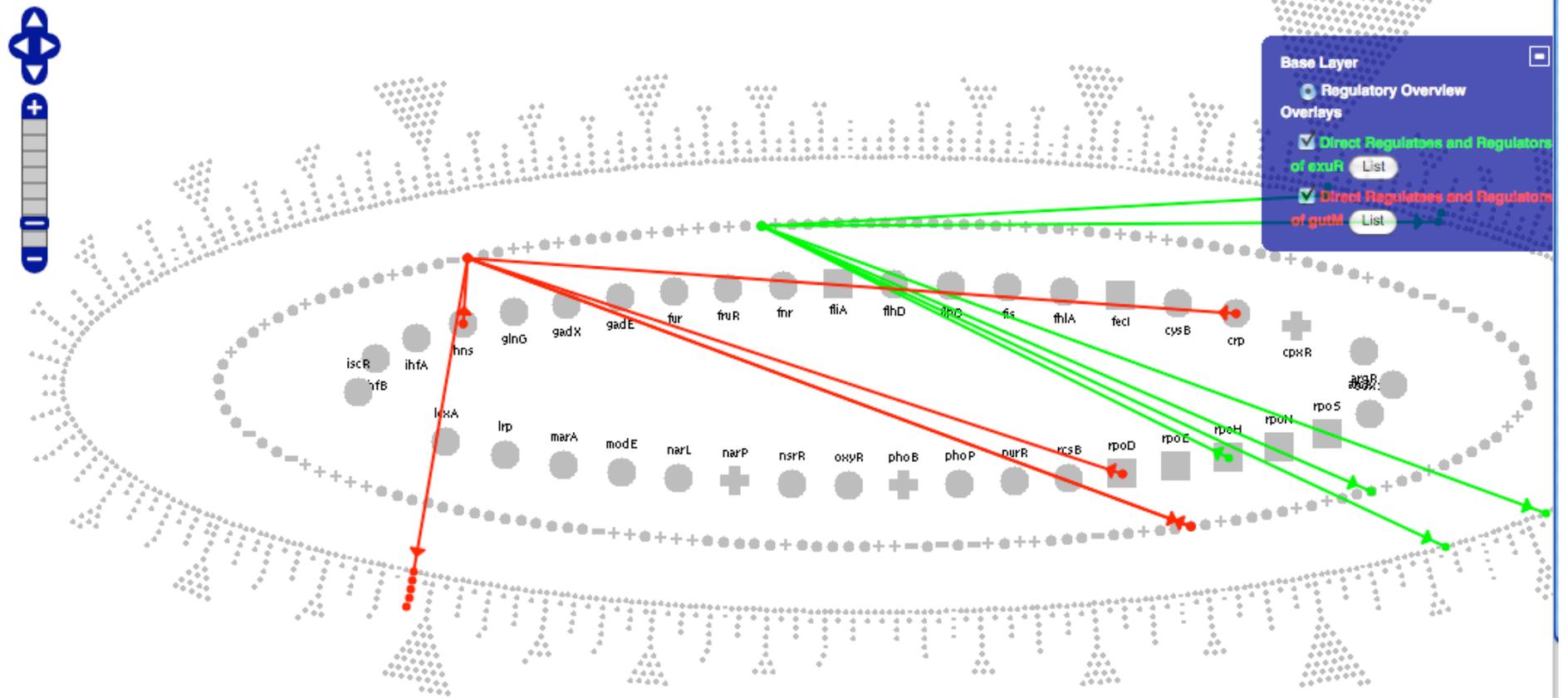
[Dismiss](#)

Base Layer
Cellular Overview
Overlays
 Compound 'pyruvate'
 Gene 'his'

DcuSR Two-Component Signal Transduction System
DpiAB Two-Component Signal Transduction System
PhoRE Two-Component Signal Transduction System
ZraSR Two-Component Signal Transduction System

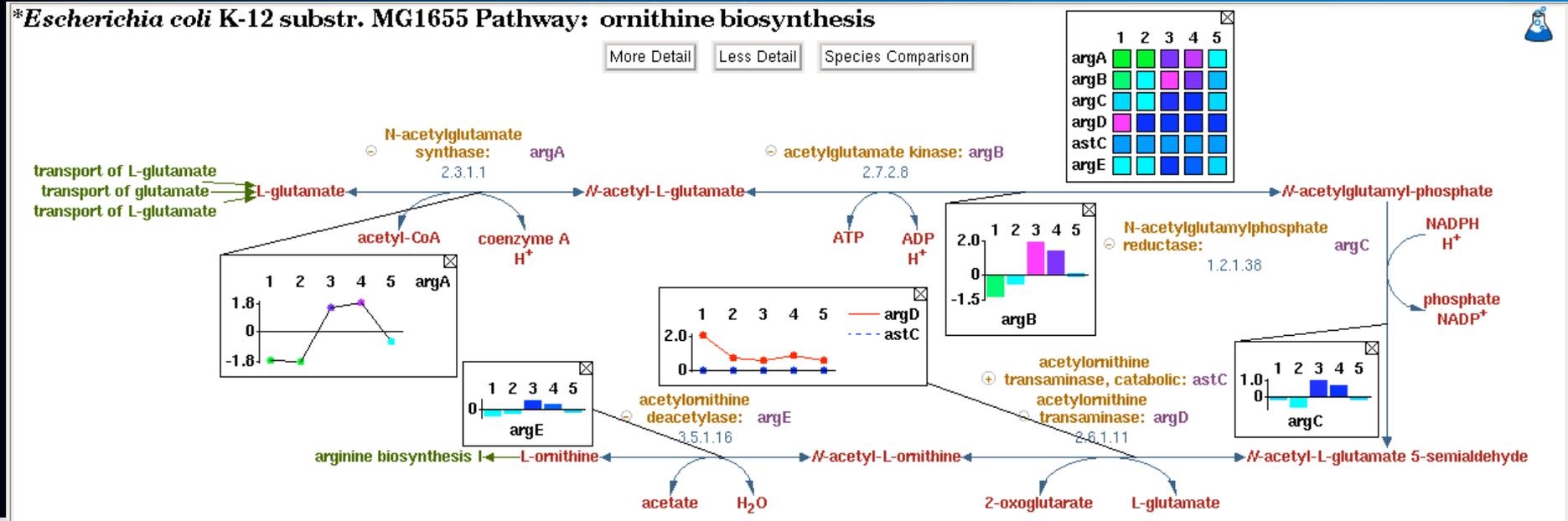
Regulatory Overview

Regulatory Overview of *Escherichia coli K-12 substr. MG1655* (Pan left/right/up/down by holding the left button mouse, left-click on gene node to open its gene page, right-click on gene node to open its regulatory overview page)

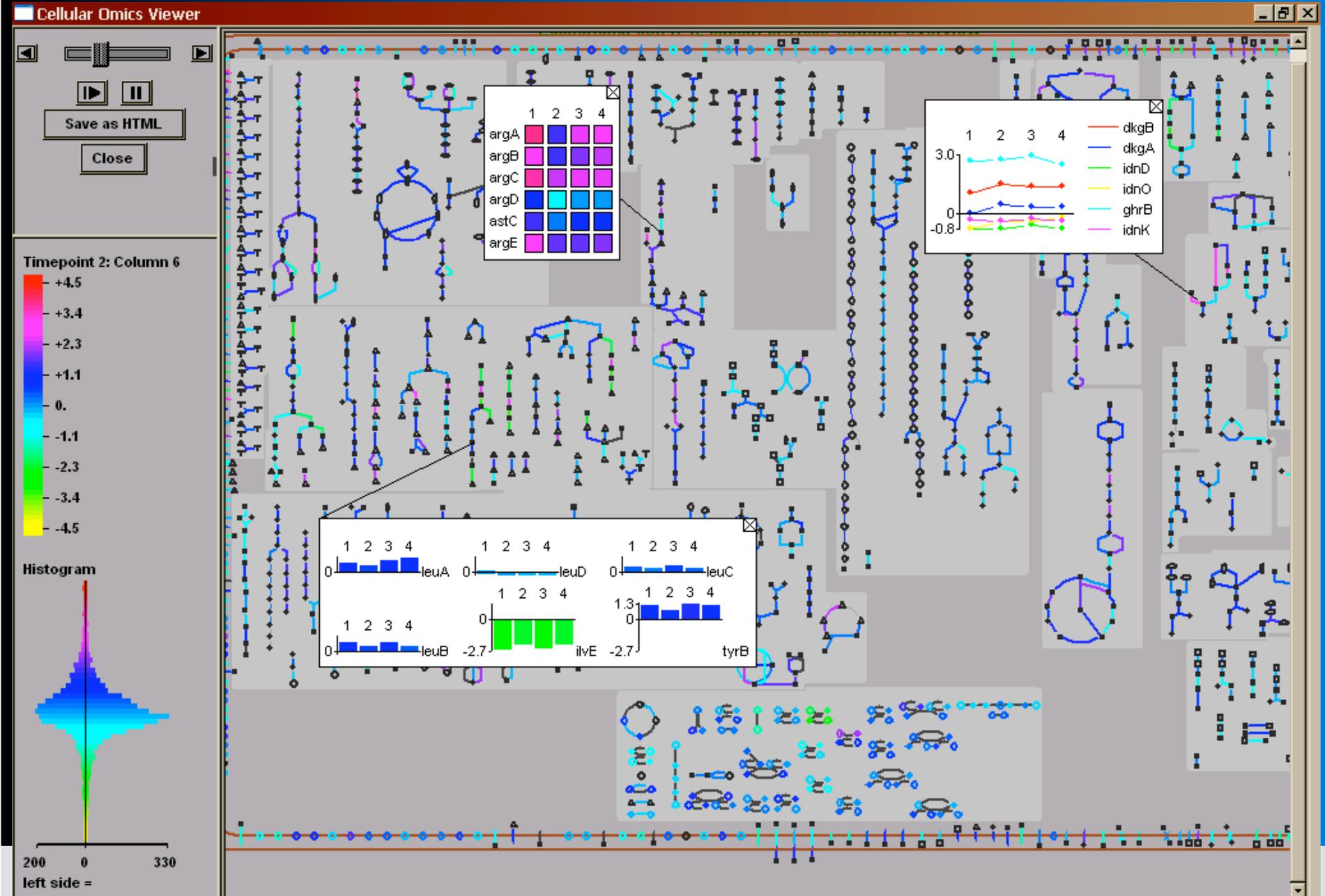


Omics Popups

- Desktop Pathway Tools only
- Can show omics popups for a gene, reaction, pathway
- Use also in Cellular Overview
- Choose from 3 styles: heatmap, bar graph, plot



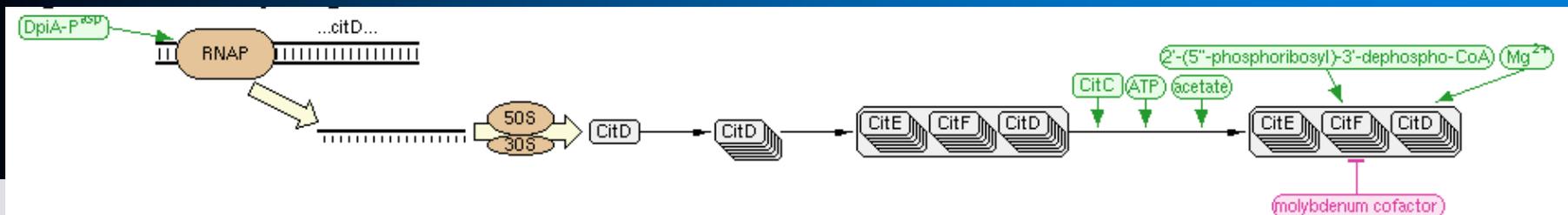
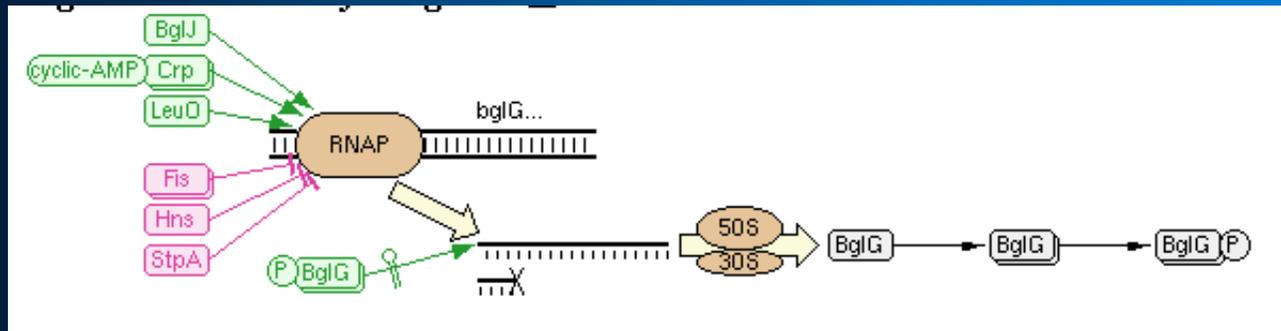
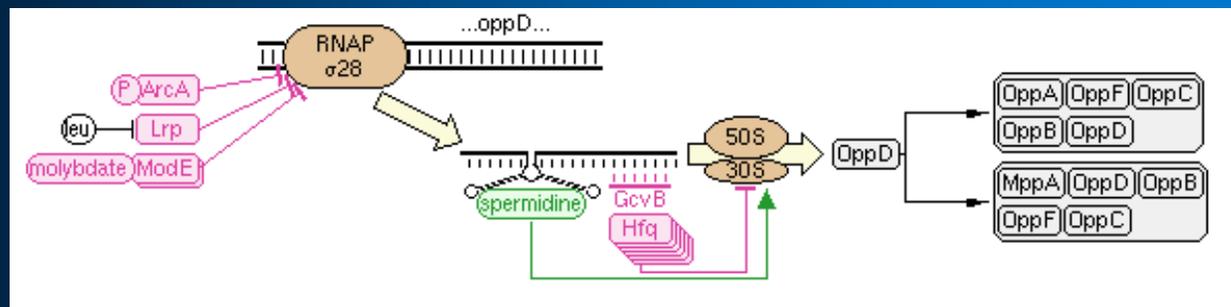
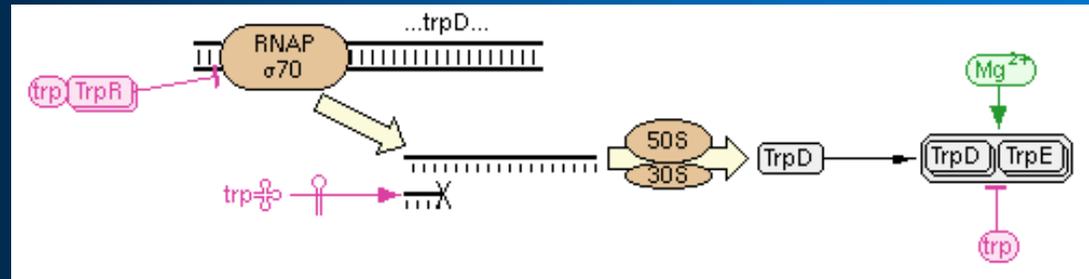
Omics Data Graphing



Pathway Tools Captures All Bacterial Regulation Mechanisms

- **Regulation of transcription**
 - By transcription factors
 - By attenuation
- **Regulation of translation**
 - By proteins and small RNAs
- **Regulation of protein activity**
 - By covalent modification (e.g., phosphorylation)
 - By non-covalent modification (e.g., allosteric inhibitors)
- **Support: Schema, editing tools, display tools**

Regulatory Summary Diagrams



Other Recent Enhancements

- **Phases I and II of upgrade to Pathway Tools Web mode**
 - Phase III still to come
- **Ability to customize pathway displays via Web site**
 - Pathway → Customize

Reachability Analysis of Metabolic Networks

- **Given:**
 - A PGDB for an organism
 - A set of initial metabolites
- **Infer:**
 - What set of products **can be** synthesized by the small-molecule metabolism of the organism
- **Motivations:**
 - Quality control for PGDBs
 - ◆ Verify that a known growth medium yields known essential compounds
 - Experiment with other growth media
 - Experiment with reaction knock-outs
- **Limitations**
 - Cannot properly handle compounds required for their own synthesis
 - Nutrients needed for reachability may be a superset of those required for growth

Romero and Karp, *Pacific Symposium on Biocomputing, 2001*

Algorithm: Forward Propagation Through Production System

- Each reaction becomes a production rule
- Each of the 21 metabolites in the nutrient set becomes an axiom

Nutrient set

Transport

Metabolite pool

Products

PGDB reaction set

“Fire” reactions



Reactants

E. coli K12 Cellular Overview



Coming Soon

- **BioCyc / EcoCyc / HumanCyc will support Web services for data retrieval**
- **iPhone app for BioCyc / EcoCyc / HumanCyc and other PGDBs**

Acknowledgements

●SRI

- Suzanne Paley, Ron Caspi, Ingrid Keseler, Carol Fulcher, Markus Krummenacker, Alex Shearer, Tomer Altman, Joe Dale, Fred Gilham, Pallavi Kaipa

●EcoCyc Collaborators

- Julio Collado-Vides, Robert Gunsalus, Ian Paulsen

●MetaCyc Collaborators

- Sue Rhee, Peifen Zhang, Kate Dreher
- Lukas Mueller, Anuradha Pujar

●Funding sources:

- NIH National Institute of General Medical Sciences
- NIH National Center for Research Resources

BioCyc.org

Learn more from BioCyc webinars: biocyc.org/webinar.shtml