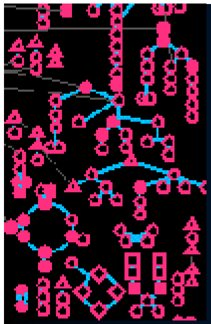


MetaCyc: A Multiorganism Database of Metabolic Pathways and Enzymes

Peter D. Karp, Ph.D.

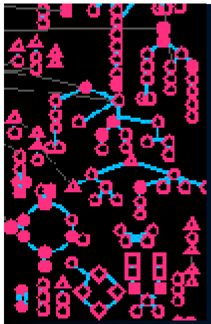
Bioinformatics Research Group

SRI International

pkarp@ai.sri.com

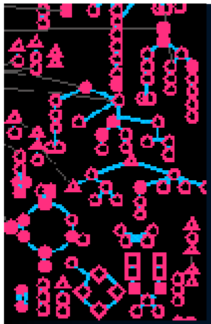
http://www.ai.sri.com/pkarp/

http://MetaCyc.org/



Overview

- **MetaCyc database**
  - Goals, content, curation strategy, applications to metabolic engineering
- **Pathway Tools software**
  - Characterize metabolic network of a sequenced organism
- **Enzyme genomics**



SRI International
Bioinformatics

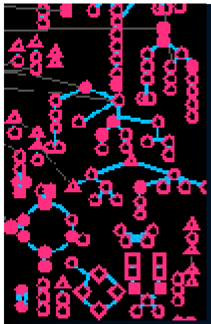
MetaCyc: Metabolic Encyclopedia
MetaCyc.org

- Nonredundant metabolic pathway database
- Describe a representative sample of every experimentally determined metabolic pathway

- Literature-based DB with extensive references and commentary
- Pathways, reactions, enzymes, substrates

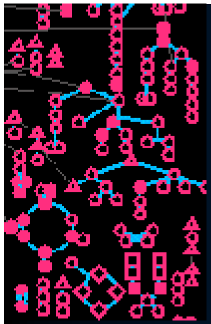
- *Nucleic Acids Research* 32:D438-442 2004.

- Jointly developed by SRI and Carnegie Institution



Applications of MetaCyc

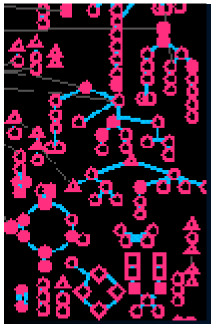
- **Reference source on metabolic pathways**
- **Metabolic engineering**
  - Find enzymes with desired activities, regulatory properties
  - Determine cofactor requirements
- **Predict pathways from genomes**
- **Systematic studies of metabolism**
- **Computer-aided education**



MetaCyc Curation

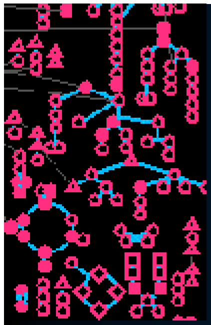
- **DB updates by 2->4 staff curators**
  - Information gathered from biomedical literature
  - Emphasis on microbial and plant pathways
  - More prevalent pathways given higher priority
  - Curator’s Guide lists curation conventions
- **Review-level database**
- **Four releases per year**

- **Quality assurance of data and software:**
  - Evaluate database consistency constraints
  - Perform element balancing of reactions
  - Display every DB object

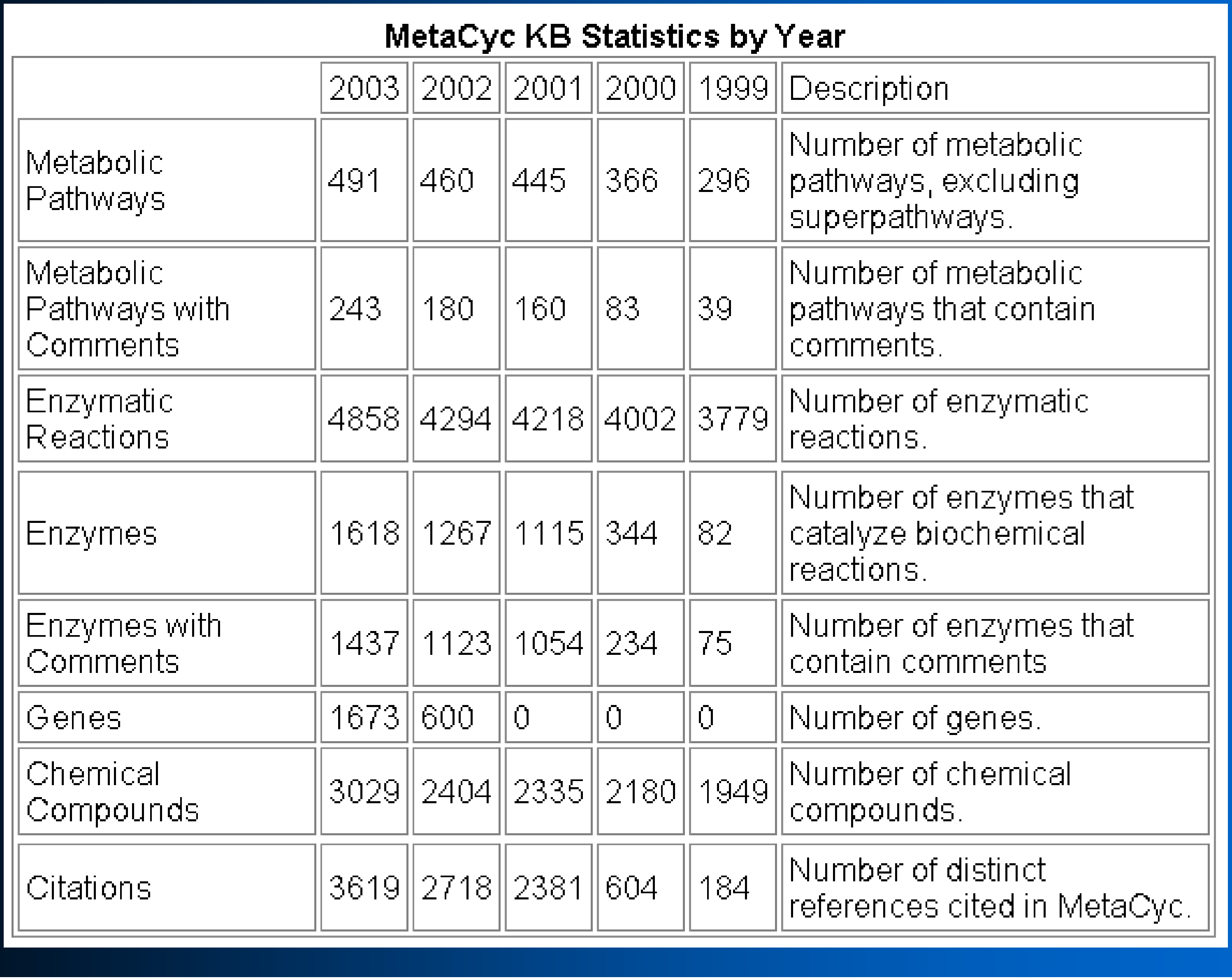


MetaCyc Curation

- **Ontologies guide querying**
  - Pathways (recently revised), compounds, enzymatic reactions
  - Example: Coenzyme M biosynthesis
- **Extensive citations and commentary**
- **Evidence codes**
  - Controlled vocabulary of evidence types
  - Attach to pathways and enzymes:
    - Code : Citation : Curator : date
- **Release notes explain recent updates**
  - http://biocyc.org/metacyc/release-notes.shtml

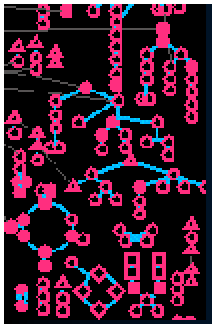


MetaCyc Data



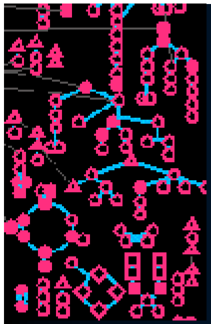
MetaCyc KB Statistics by Year

	2003	2002	2001	2000	1999	Description
Metabolic Pathways	491	460	445	366	296	Number of metabolic pathways, excluding superpathways.
Metabolic Pathways with Comments	243	180	160	83	39	Number of metabolic pathways that contain comments.
Enzymatic Reactions	4858	4294	4218	4002	3779	Number of enzymatic reactions.
Enzymes	1618	1267	1115	344	82	Number of enzymes that catalyze biochemical reactions.
Enzymes with Comments	1437	1123	1054	234	75	Number of enzymes that contain comments
Genes	1673	600	0	0	0	Number of genes.
Chemical Compounds	3029	2404	2335	2180	1949	Number of chemical compounds.
Citations	3619	2718	2381	604	184	Number of distinct references cited in MetaCyc.



MetaCyc Frequent Organisms

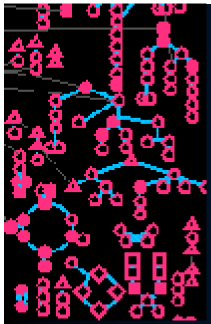
Escherichia coli	156
Arabidopsis thaliana	47
Homo sapiens	30
Salmonella typhimurium	20
Bacillus subtilis	20
Sulfolobus solfataricus	18
Pseudomonas putida	14
Saccharomyces cereivisae	14
Haemophilus influenzae	13
Glycine max	11
Deinococcus radiodurans	10
Mycoplasma capricolum	9



MetaCyc Data

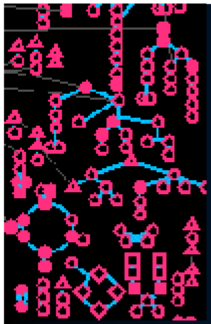
- **Of the 1548 enzymes:**
  - 818 are monomers
  - 730 are multimers
  - 570 are homomultimers, 160 are heteromultimers
- **Enzymes with cofactors: 512**
- **Enzymes with activators or inhibitors: 577**

- **Average pathway length: 5 reactions**



MetaCyc Pathway Variants

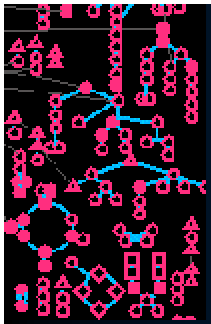
- **Pathways that accomplish similar biochemical functions using different biochemical routes**
  - Alanine biosynthesis I – *E. coli*
  - Alanine biosynthesis II – *H. sapiens*
- **Pathways that accomplish similar biochemical functions using similar sets of reactions**
  - Several variants of TCA Cycle



MetaCyc Super-Pathways

- **Groups of pathways linked by common substrates**
- **Example: Super-pathway containing**
  - Chorismate biosynthesis
  - Tryptophan biosynthesis
  - Phenylalanine biosynthesis
  - Tyrosine biosynthesis

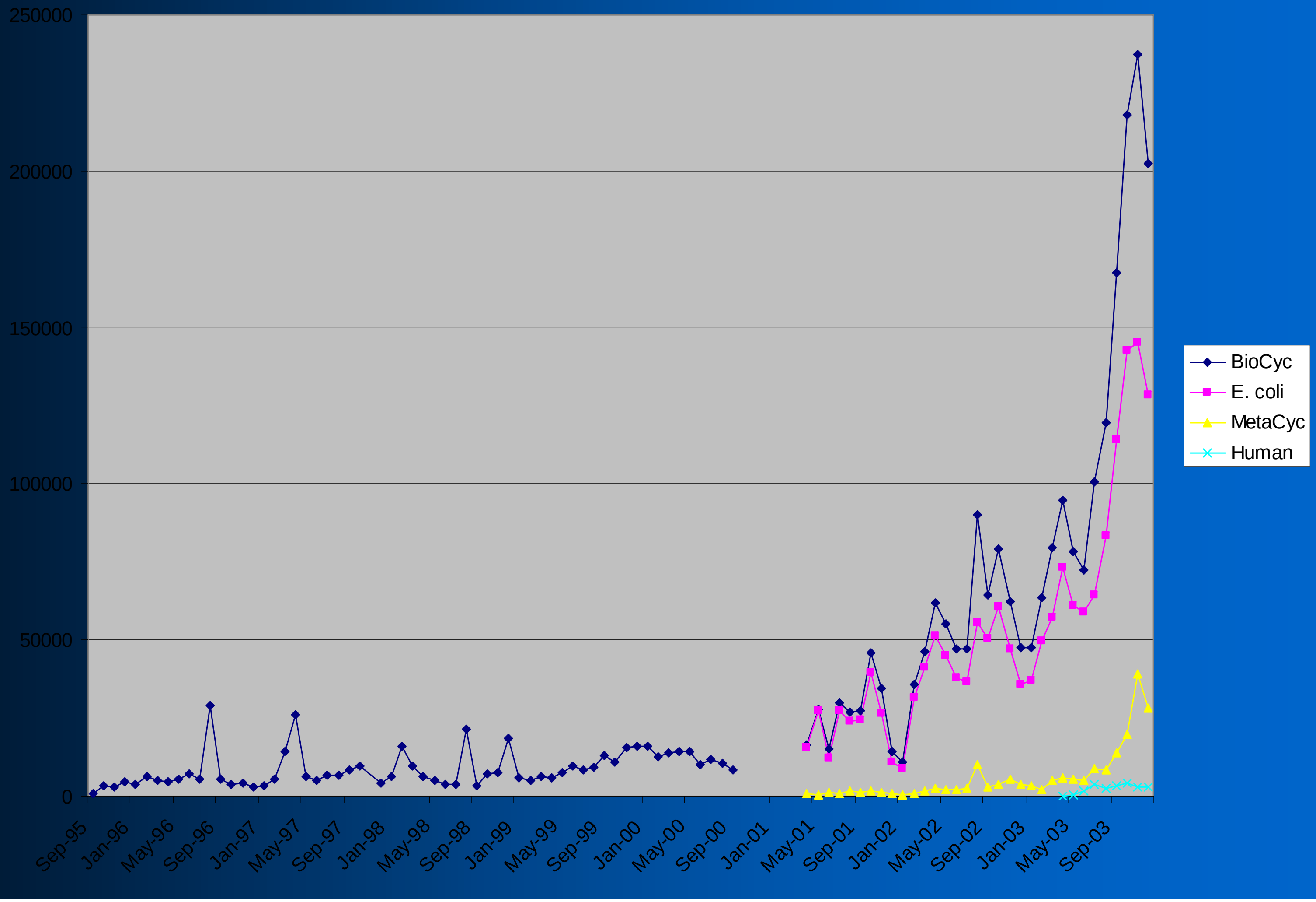
- **Super-pathways defined by listing their component pathways**
- **Multiple levels of super-pathways can be defined**
- **Pathway layout algorithms accommodate super-pathways**

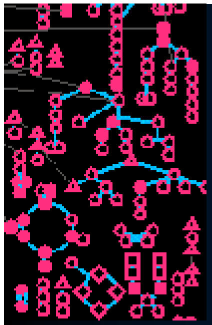


EcoCyc Compared to Overlapping Databases

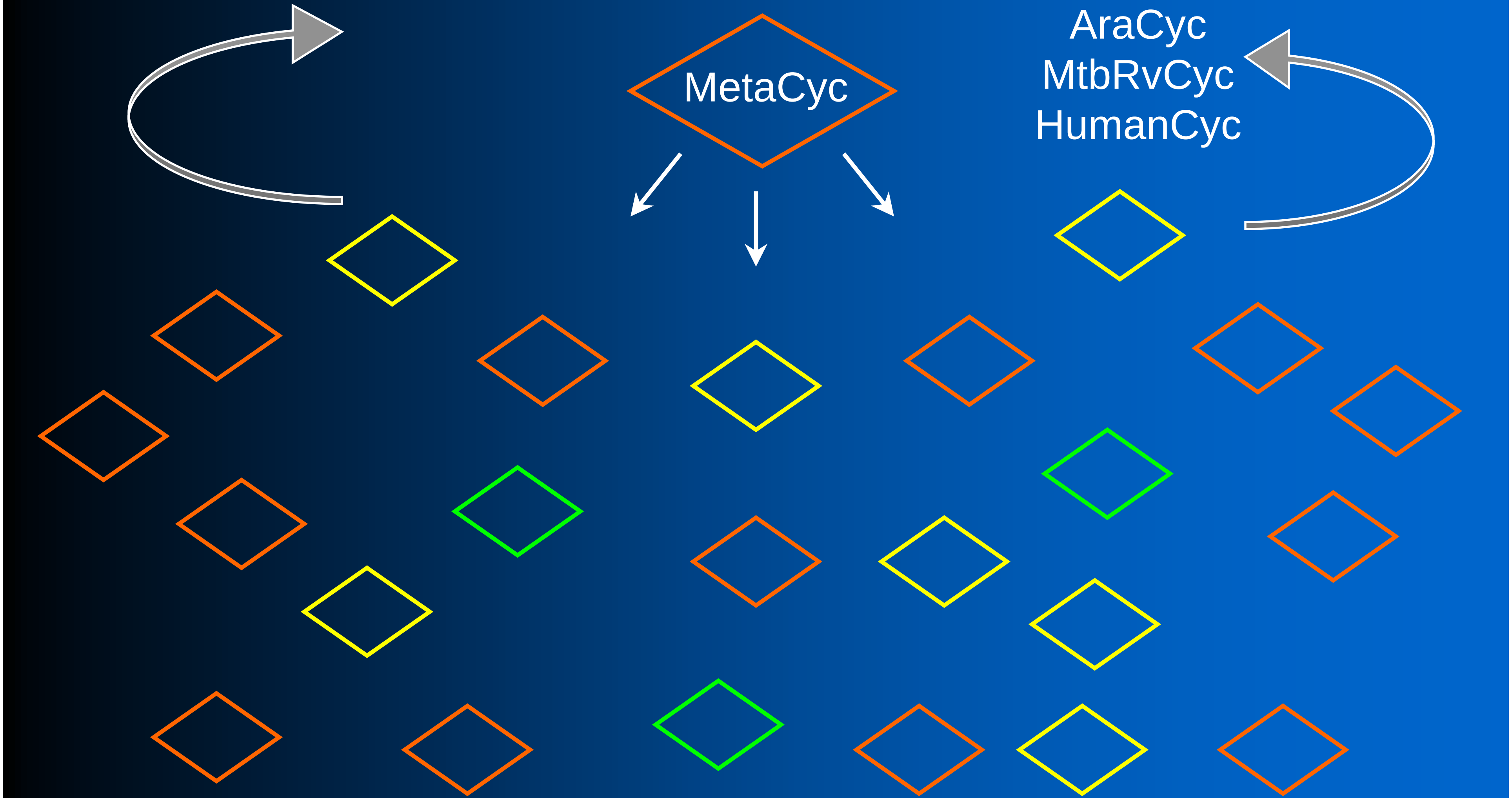
SRI International Bioinformatics

- Downloads of software/database bundle: 251
- Web site traffic: (hits per month)

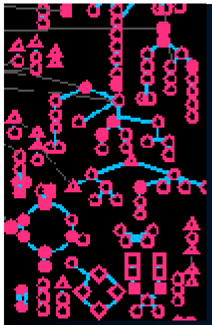




Family of Pathway/Genome Databases

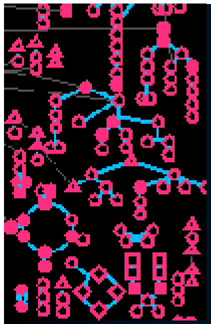


EcoCyc
CauloCyc
AraCyc
MtbRvCyc
HumanCyc



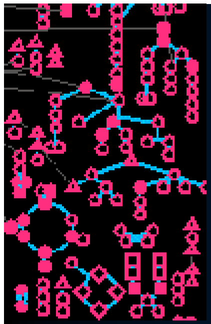
MetaCyc Advisory Board

- **Dale Kaiser, Stanford**
- **Patsy Babbitt, UCSF**
- **Mark Stitt**
- **Trey Ideker, UC San Diego**
- **Chris Somerville, Carnegie Institution**
- **Jay Keasling, UC Berkeley**
- **Jean-Francois Tomb, Dupont**
- **Fernando Valle, Genencor**
- **Russ Altman, Stanford**



Comparison of BioCyc to KEGG

- **KEGG approach: Static collection of pathway diagrams that are color-coded to produce organism-specific views**
- **KEGG vs MetaCyc: Resource on literature-derived pathways**
  - KEGG pathways maps are composites of pathways in many organisms -- do not identify what specific pathways elucidated in what organisms
  - KEGG has no literature citations, no comments, less enzyme detail
- **KEGG vs BioCyc organism-specific PGDBs**
  - KEGG covers more organisms than does BioCyc
  - KEGG does not curate or customize pathway networks for each organism
- **Software tools**
  - KEGG has no algorithmic visualization tools
  - KEGG has no queryable metabolic-map overview diagram
  - KEGG has no interactive editing tools

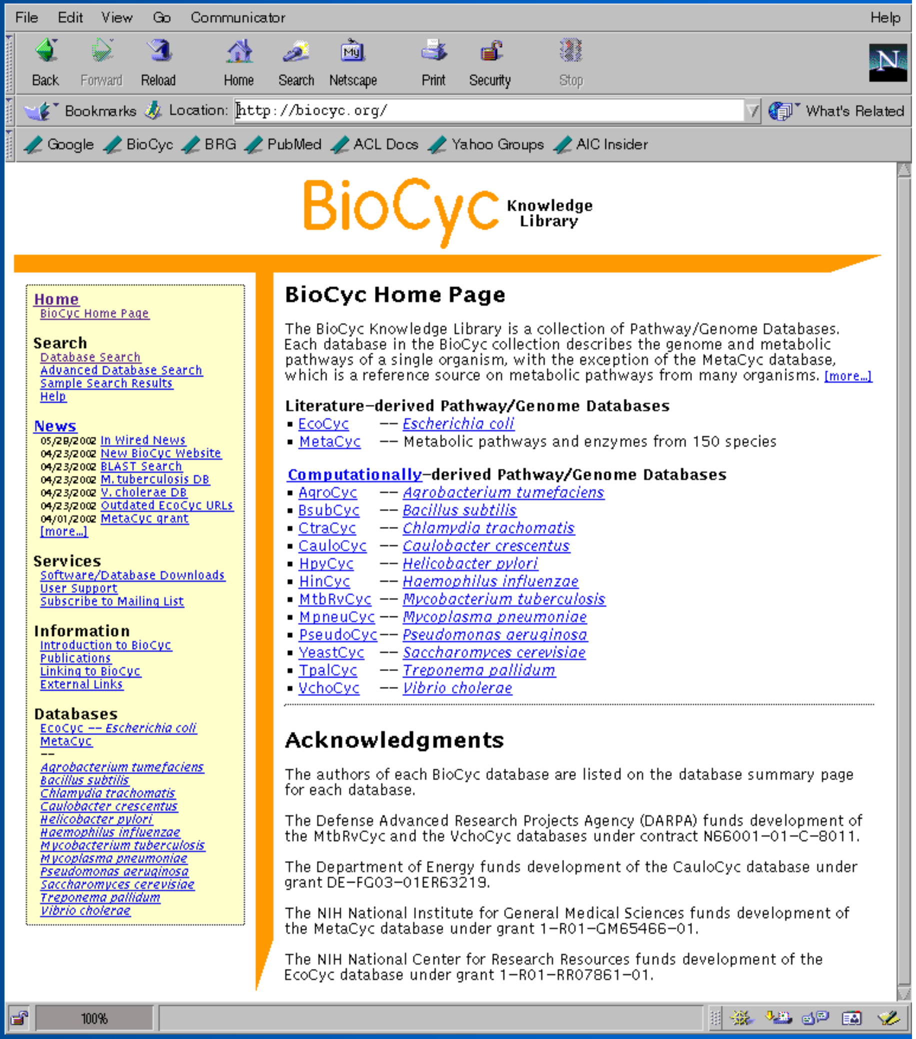


Terminology

- **Model Organism Database (MOD) – DB describing genome and other information about an organism**

- **Pathway/Genome Database (PGDB) – MOD that combines information about**
    - Pathways, reactions, substrates
    - Enzymes, transporters
    - Genes, replicons
    - Transcription factors, promoters, operons, DNA binding sites

- **BioCyc – Collection of 15 PGDBs at BioCyc.org**
    - EcoCyc, AgroCyc, YeastCyc



Home
BioCyc Home Page

Search
Database Search
Advanced Database Search
Sample Search Results
Help

News
05/28/2002 In Wired News
04/23/2002 New BioCyc Website
04/23/2002 BLAST Search
04/23/2002 M. tuberculosis DB
04/23/2002 V. cholerae DB
04/23/2002 Outdated EcoCyc URLs
04/01/2002 MetaCyc grant
[more...]

Services
Software/Database Downloads
User Support
Subscribe to Mailing List

Information
Introduction to BioCyc
Publications
Linking to BioCyc
External Links

Databases
EcoCyc -- *Escherichia coli*
MetaCyc
--
Agrobacterium tumefaciens
Bacillus subtilis
Chlamydia trachomatis
Caulobacter crescentus
Helicobacter pylori
Haemophilus influenzae
Mycobacterium tuberculosis
Mycoplasma pneumoniae
Pseudomonas aeruginosa
Saccharomyces cerevisiae
Treponema pallidum
Vibrio cholerae

BioCyc Home Page

The BioCyc Knowledge Library is a collection of Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism, with the exception of the MetaCyc database, which is a reference source on metabolic pathways from many organisms. [more...]

Literature-derived Pathway/Genome Databases

- EcoCyc -- *Escherichia coli*
- MetaCyc -- Metabolic pathways and enzymes from 150 species

Computationally-derived Pathway/Genome Databases

- AgroCyc -- *Agrobacterium tumefaciens*
- BsubCyc -- *Bacillus subtilis*
- CtraCyc -- *Chlamydia trachomatis*
- CauloCyc -- *Caulobacter crescentus*
- HpyCyc -- *Helicobacter pylori*
- HinCyc -- *Haemophilus influenzae*
- MtbRvCyc -- *Mycobacterium tuberculosis*
- MpneuCyc -- *Mycoplasma pneumoniae*
- PseudoCyc -- *Pseudomonas aeruginosa*
- YeastCyc -- *Saccharomyces cerevisiae*
- TpalCyc -- *Treponema pallidum*
- VchoCyc -- *Vibrio cholerae*

Acknowledgments

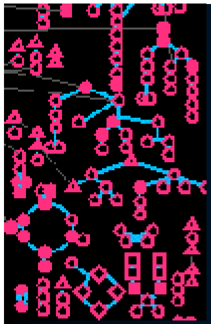
The authors of each BioCyc database are listed on the database summary page for each database.

The Defense Advanced Research Projects Agency (DARPA) funds development of the MtbRvCyc and the VchoCyc databases under contract N66001-01-C-8011.

The Department of Energy funds development of the CauloCyc database under grant DE-FG03-01ER63219.

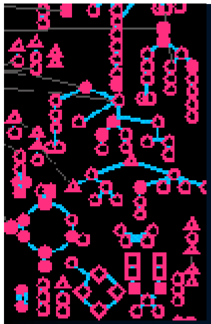
The NIH National Institute for General Medical Sciences funds development of the MetaCyc database under grant 1-R01-GM65466-01.

The NIH National Center for Research Resources funds development of the EcoCyc database under grant 1-R01-RR07861-01.



Pathway Tools Software

- **PathoLogic**
  - Computational creation of new Pathway/Genome Databases
  - Predict metabolic network, operons, pathway hole fillers
- **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
- **Bioinformatics 18:S225 2002**

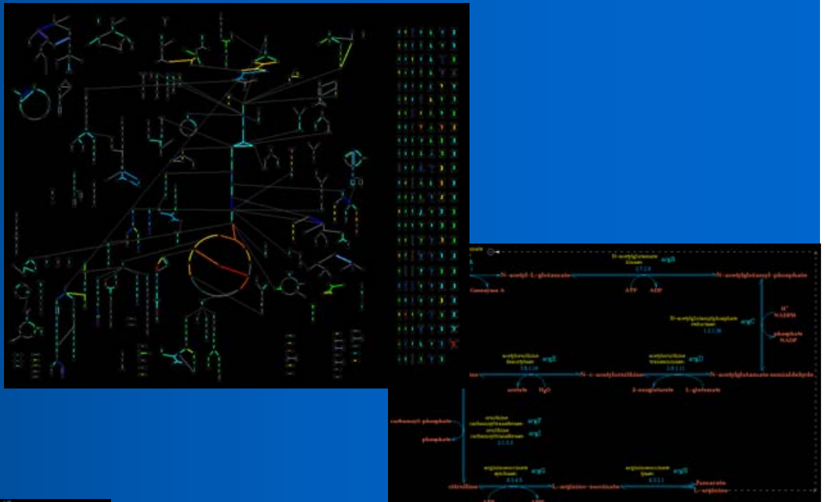


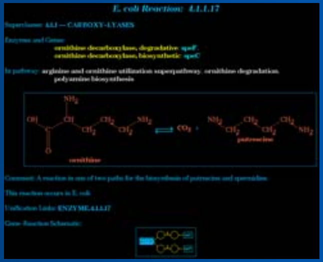
Pathway Tools Algorithms

SRI International Bioinformatics

- **Visualization and editing tools for following datatypes**

- **Full Metabolic Map**
  - Paint gene expression data on metabolic network; compare metabolic networks
- **Pathways**
  - Pathway prediction
- **Reactions**
  - Balance checker
- **Compounds**
  - Chemical substructure comparison
- **Enzymes, Transporters, Transcription Factors**
- **Genes**
- **Chromosomes**
- **Operons**
  - Operon prediction





E. coli Reaction: 4.1.1.17

Superclass: 4.1.1 — CARBOXY-LYASES

Enzymes and Genes:
ornithine decarboxylase, degradative: speF
ornithine decarboxylase, biosynthetic: speC

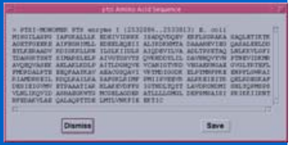
In pathway: arginine and ornithine utilization superpathway, ornithine degradation, polyamine biosynthesis

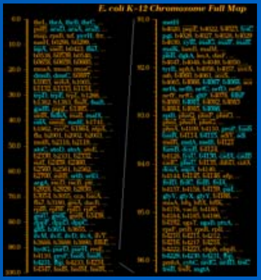
Comment: A reaction in one of two paths for the biosynthesis of putrescine and spermidine.

This reaction occurs in E. coli

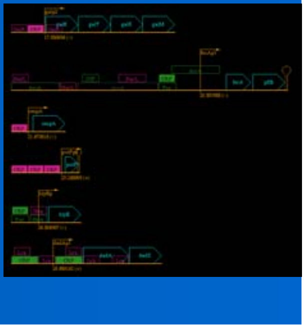
Unification Links: ENZYME:4.1.1.17

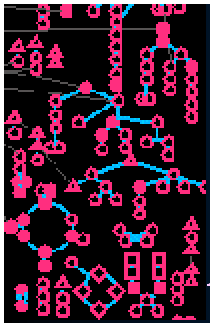
Gene-Reaction Schematic:



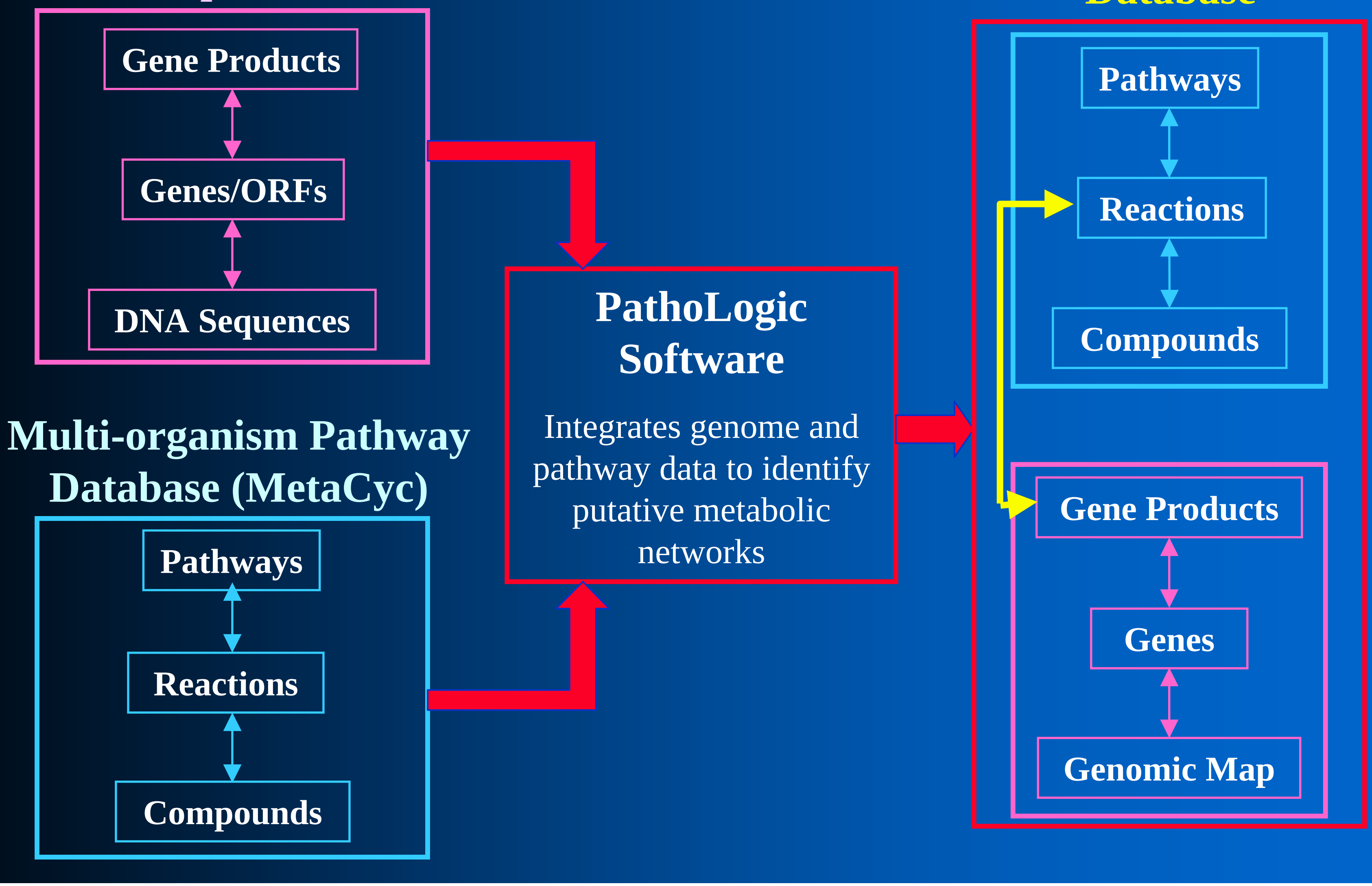


E. coli K-12 Chromosome Full Map





Inference of Metabolic Pathways



Annotated Genomic Sequence

Gene Products

Genes/ORFs

DNA Sequences

Multi-organism Pathway Database (MetaCyc)

Pathways

Reactions

Compounds

PathoLogic Software

Integrates genome and pathway data to identify putative metabolic networks

Pathway/Genome Database

Pathways

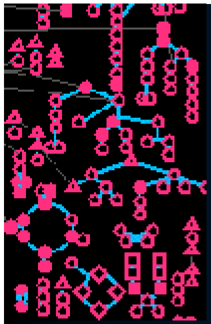
Reactions

Compounds

Gene Products

Genes

Genomic Map



BioCyc Collection of Pathway/Genome DBs

SRI International
Bioinformatics

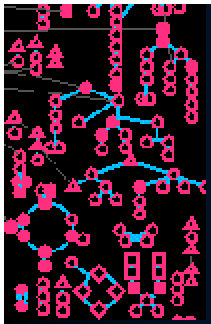
- **Literature-based Datasets:**
- ***MetaCyc***
- ***Escherichia coli K-12 -- (EcoCyc)***

http://BioCyc.org/

225,00 hits/month

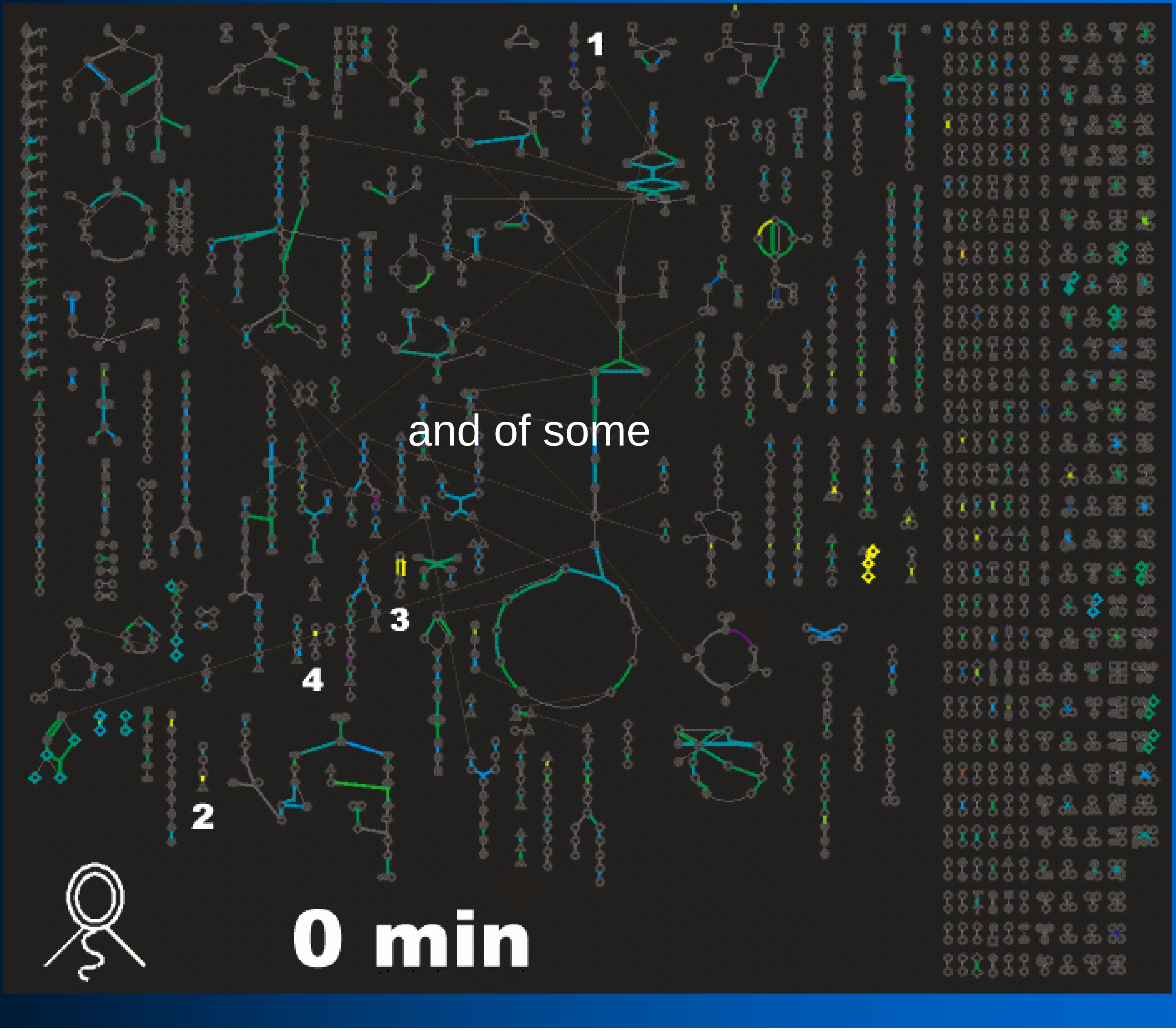
Computationally Derived Datasets:

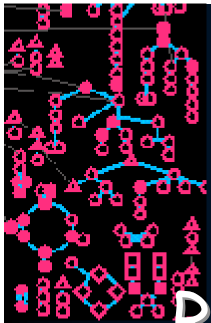
- ***Homo sapiens***
- ***Agrobacterium tumefaciens***
- ***Caulobacter crescentus***
- ***Chlamydia trachomatis***
- ***Bacillus subtilis***
- ***Escherichia coli O157:H7***
- ***Helicobacter pylori***
- ***Haemophilus influenzae***
- ***Mycobacterium tuberculosis RvH37***
- ***Mycobacterium tuberculosis CDC1551***
- ***Mycoplasma pneumonia***
- ***Pseudomonas aeruginosa***
- ***Shigella flexneri***
- ***Treponema pallidum***
- ***Vibrio cholerae***



SRI International
Bioinformatics

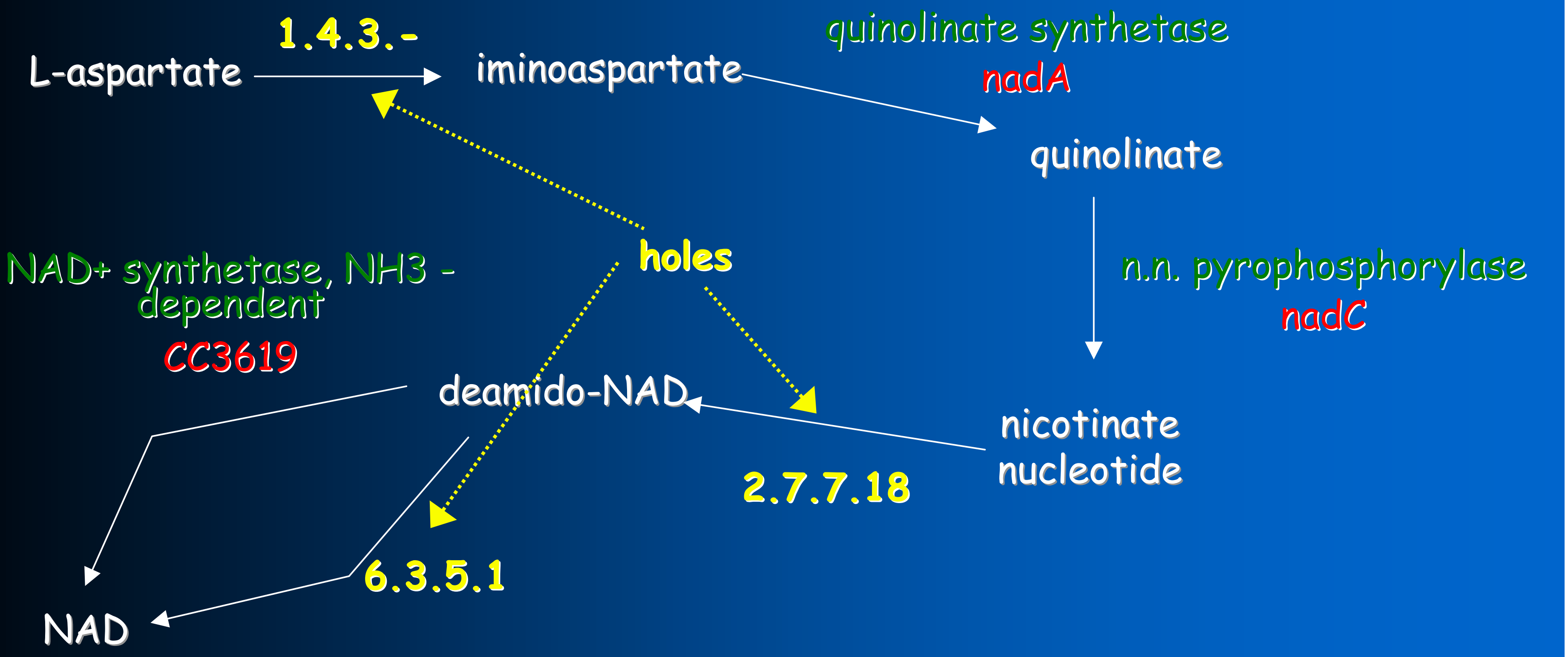
C. crescentus Cell Cycle Gene Expression





Filling Pathway Holes

Reactions in the pathway without enzymes assigned are pathway holes



L-aspartate

1.4.3.-

iminoaspartate

quinolinate synthetase

nadA

quinolinate

n.n. pyrophosphorylase

nadC

nicotinate nucleotide

2.7.7.18

deamido-NAD

holes

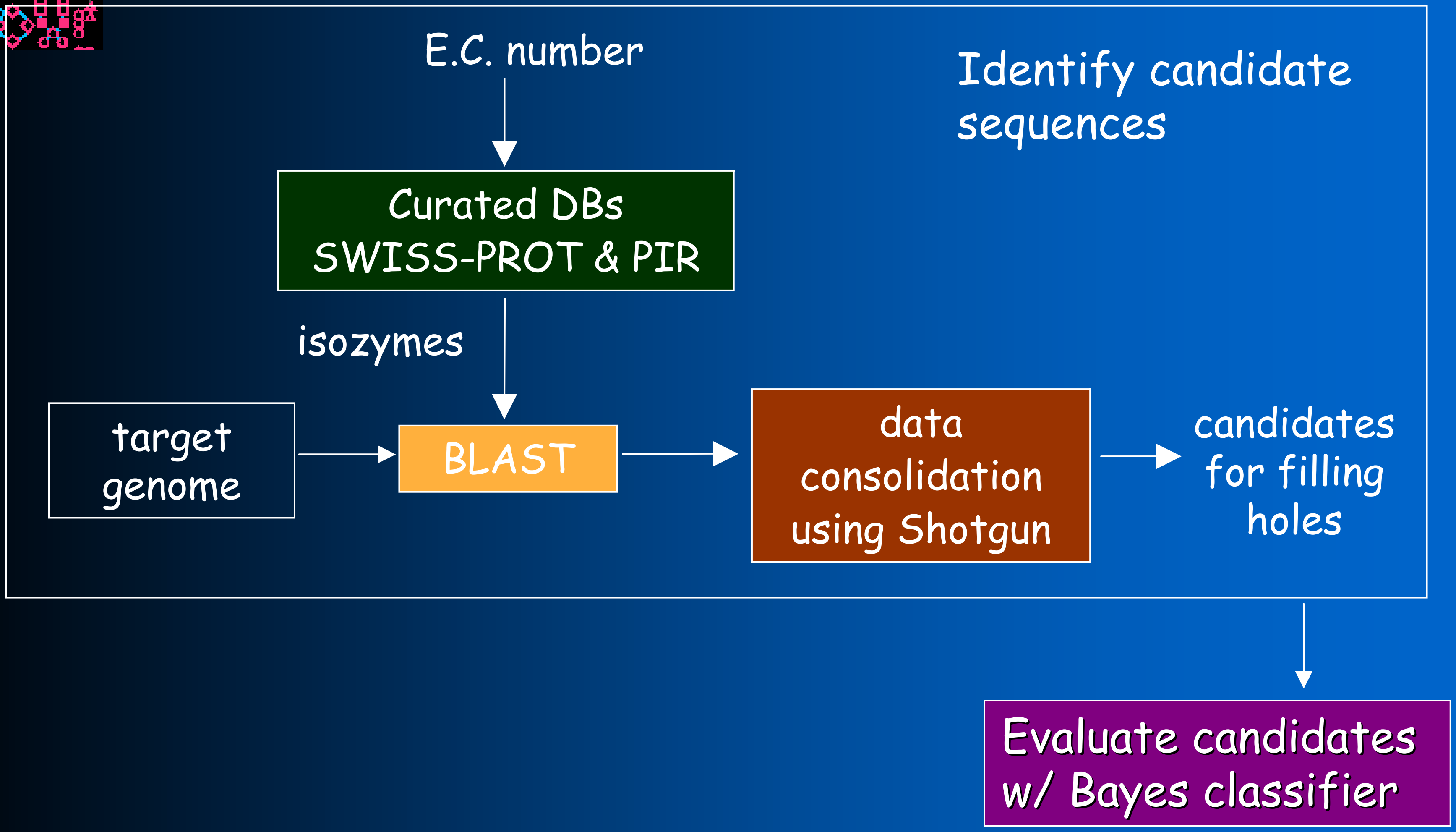
NAD+ synthetase, NH3 - dependent

CC3619

6.3.5.1

NAD

Approach for filling pathway holes



Identify candidate sequences

E.C. number

Curated DBs
SWISS-PROT & PIR

isozymes

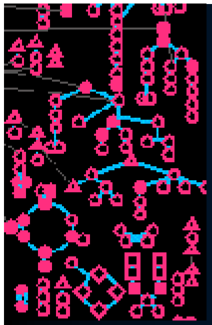
target genome

BLAST

data consolidation using Shotgun

candidates for filling holes

Evaluate candidates w/ Bayes classifier



CauloCyc Pathway Holes

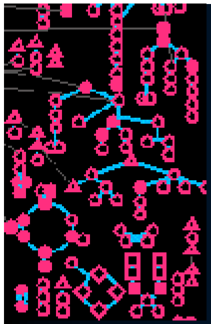
SRI International
Bioinformatics

Fill holes by predicting the probability that a gene has a particular function

- 130 pathways containing 582 reactions
- 92 pathways w/ at least 1 missing reaction
- 236 missing reactions

CauloCyc holes filled:

- 77 holes filled at P >0.9
- 4 ORFs identified at P >0.5
  - (3 ORFs, P >= 0.9)
- Multifunctional enzymes
- Enzymes with different functional assignments
- Enzymes with imprecise functional assignments



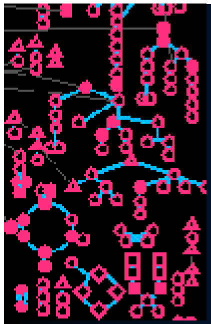
Pathway/Genome DBs Created by External Users

- ***Saccharomyces cerevisiae, Stanford University***
  - pathway.yeastgenome.org/biocyc/
- ***Plasmodium falciparum*, Stanford University**
  - plasmocyc.stanford.edu
- ***Mycobacterium tuberculosis*, Stanford University**
  - BioCyc.org

- ***Arabidopsis thaliana* and *Synechocystis*, Carnegie Institution of Washington**
  - Arabidopsis.org:1555

- ***Methanococcus janaschii*, European Bioinformatics Institute**
  - Maine.ebi.ac.uk:1555

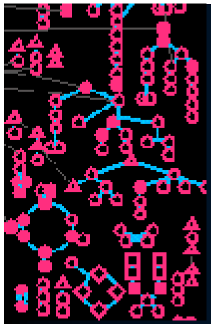
- **40 PGDBs created; 20 more in progress**
- **Software freely available -- 70 licensed users**
- **Each PGDB owned by its creator**



Biochemically Characterized Enzymes with No Known Sequence

- **1.1.3.40**
  - D-mannitol oxidase
  - mannitol + O_2 = mannose + H_2O_2
  - Vorhaben et al, 1986; isolated from snail digestive gland tissue

- **1.1.1.7**
  - propanediol-phosphate dehydrogenase
  - propane-1,2-diol 1-phosphate + NAD+ = hydroxyacetone phosphate + NADH + H+
  - Sellinger and Miller, 1959



Unsequenced Enzymes

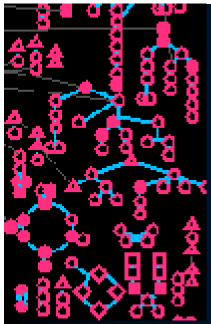
- ENZYME DB lists 4208 EC#s (v33.0)

- Swiss-Prot (version 42.6) references 1791 distinct EC#s
- TrEMBL (version 25.4) references 300 more EC#s
- PIR (PIR-PSD version 78.03) references 158 more EC#s

- CMR (version April-2003) references 23 more EC#s
- BioCyc (version 7.6) references 57 more EC#s

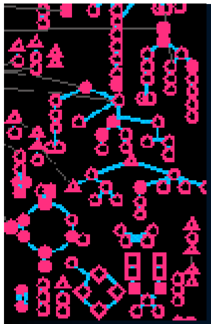
- These databases reference 2329 distinct EC#s, or 55% of all known EC numbers

- Therefore, for 1879 (=4208-2329) EC numbers (45%), no sequence is known



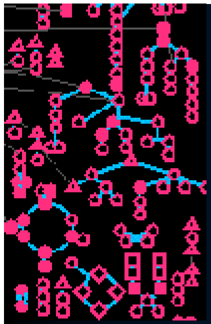
BioWarehouse

- **SRI bioinformatics database integration platform**
- **Loader bioinformatics databases into relational database warehouse**
  - Oracle and MySQL implementations
- **Databases supported:**
  - SwissProt, TrEMBL
  - BioCyc, KEGG, ENZYME
  - CMR
  - NCBI-Taxonomy



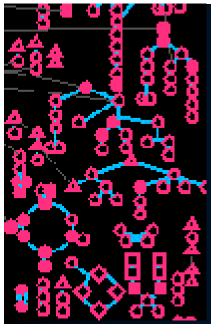
Caveats

- **The universe of enzyme activities is larger than 4200**
- **It could be that many more enzymes are sequenced, but have not been assigned EC#s in protein DBs**
  - SwissProt: 91,000 -> 500 -> 50 -> 4



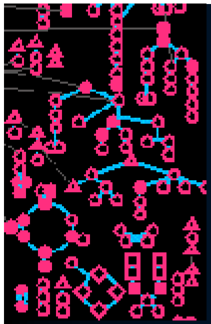
Enzyme Genomics Initiative

- **Unsequenced enzymes:**
  - Cannot be recognized in sequenced genomes
  - Decrease accuracy of metabolic pathway prediction
  - Cannot be subjects of metabolic engineering
- **Attempt to assign a sequence to all known enzymatic functions**
- **Analogous to structural genomics initiative**
- **Example method:**
  - Organism from which enzyme was purified has been sequenced
  - Computationally match biochemical properties of enzyme to all ORFs



Pathway Exchange Format

- **BioPAX = Biopathway Exchange Language**
- **A data exchange format intended to facilitate sharing of pathway data**
- **BioPAX will provide a consistent format for pathway data so it will be easier for consumers of pathway data (e.g. tool developers, DB curators) to integrate data from multiple sources**
- **Approach:**
  - Study datatype definitions from multiple pathway databases
  - Group discussion of common and idiosyncratic elements
  - XML/OWL based definition of these datatypes

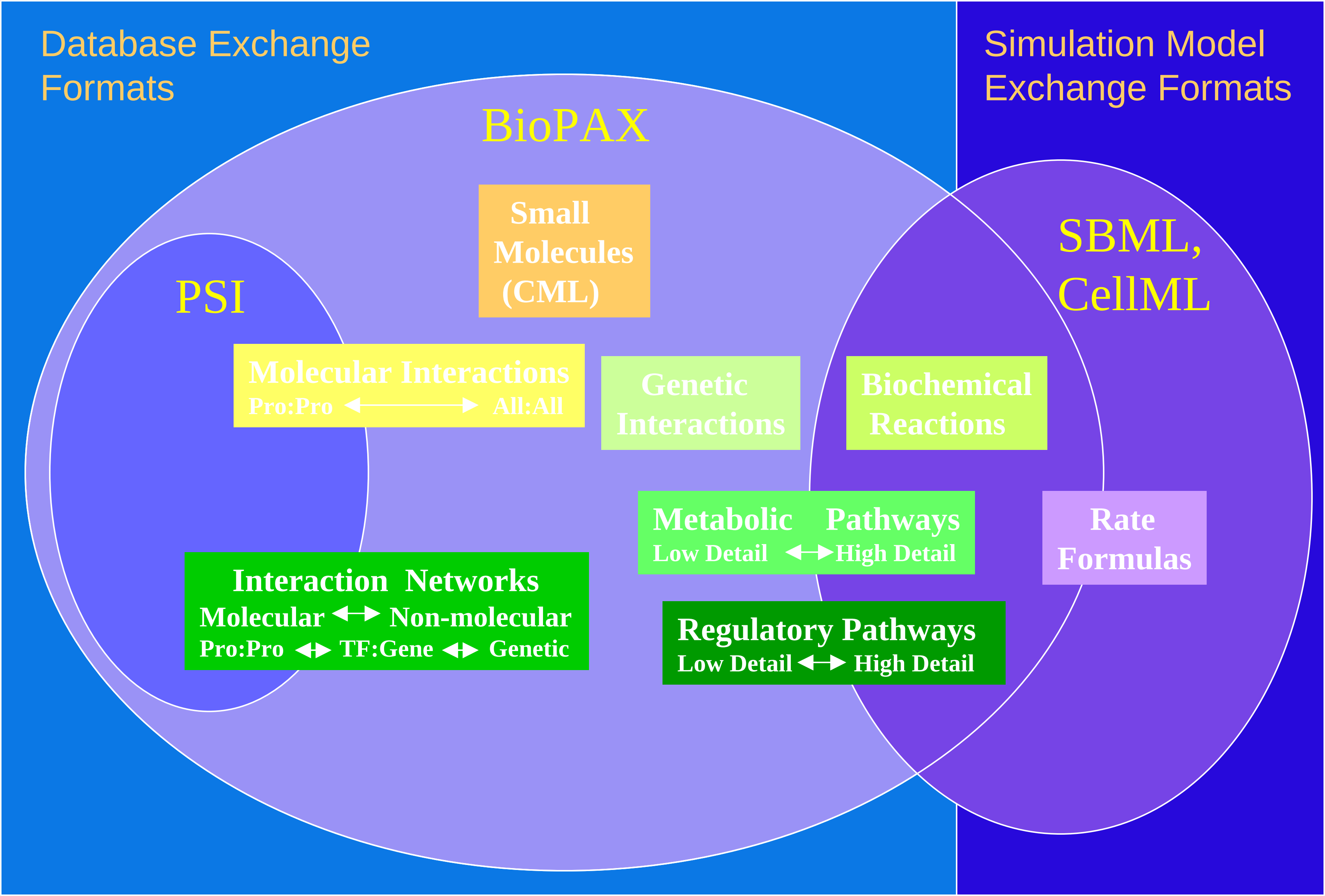


BioPAX Roadmap

- **Level 1 – Due in May 2004**
  - Datatypes: Small molecules, Proteins, RNAs, Biochemical reactions, Enzyme catalysis, Complex Assembly
  - Data source compatibility: BioCyc, WIT, KEGG, Amaze, GK
  - Deliver OWL definitions, specification document, translators for several DBs
- **Level 2 – Binding Interactions**
- **Level 3 – Genetic interactions, Gene Regulation**
- **Future levels – Signal transduction**

SRI International
Bioinformatics

Exchange Formats in the Pathway Data Space



Database Exchange Formats

Simulation Model Exchange Formats

BioPAX

PSI

SBML, CellML

Small Molecules (CML)

Molecular Interactions
Pro:Pro ←→ All:All

Genetic Interactions

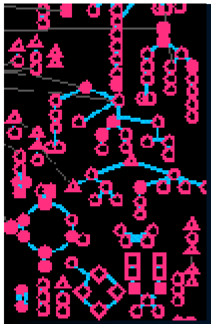
Biochemical Reactions

Metabolic Pathways
Low Detail ←→ High Detail

Rate Formulas

Interaction Networks
Molecular ←→ Non-molecular
Pro:Pro ←→ TF:Gene ←→ Genetic

Regulatory Pathways
Low Detail ←→ High Detail



BioPAX Supporting Groups

Groups

- Memorial Sloan-Kettering Cancer Center**: C. Sander, J. Luciano, M. Cary, G. Bader**
- University of Colorado Health Sciences Center**: I. Shah**
- SRI Bioinformatics Research Group**: P. Karp, S. Paley, J. Pick**
- BioPathways Consortium**: J. Luciano, Eric Neumann, Vincent Schachter (www.biopathways.org)**
- Argonne National Laboratory**: N. Maltsev**
- Samuel Lunenfeld Research Institute**: C. Hogue**
- Harvard CGR**: Aviv Regev**

Collaborating Organizations:

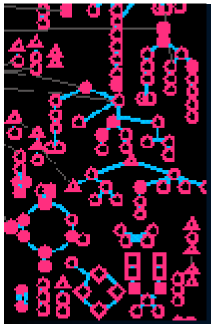
- Proteomics Standards Initiative **(psidev.sf.net)**
- Chemical Markup Language **(www.xml-cml.org)**
- SBML **(www.sbml.org)**
- CellML **(www.cellml.org)**

Databases

- BioCyc **(www.biocyc.org)**
- BIND **(www.bind.ca)**
- WIT **(wit.mcs.anl.gov/WIT2)**

Grants

- Department of Energy

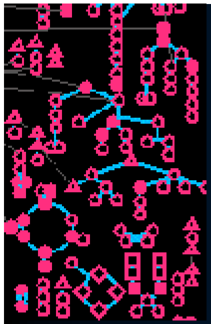


MetaCyc and Pathway Tools Availability

- **WWW MetaCyc: MetaCyc.org**

- **MetaCyc downloads freely available to non-profits**
  - Flatfiles downloadable from BioCyc.org
  - Binary executable:
    - Sun UltraSparc-170 w/ 64MB memory
    - PC, 400MHz CPU, 64MB memory, Windows or Linux
  - PerlCyc and JavaCyc APIs

- **Pathway Tools freely available to non-profits**



SRI International Bioinformatics

Acknowledgements

- **SRI**
  - Suzanne Paley, Pedro Romero, John Pick, Cindy Krieger, Martha Arnaud, Randy Gobbel, Michelle Green
- **EcoCyc Project**
  - J. Collado-Vides, J. Ingraham, I. Paulsen, M. Saier
- **MetaCyc Project**
  - Sue Rhee, Lukas Mueller, Peifen Zhang, Chris Somerville
- **Stanford**
  - Gary Schoolnik, Harley McAdams, Lucy Shapiro, Russ Altman, Iwei Yeh
- **Funding sources:**
  - NIH National Institute of General Medical Sciences
  - NIH National Human Genome Research Institute
  - NIH National Center for Research Resources
  - Department of Energy Microbial Cell Project
  - DARPA BioSpice, UPC

BioCyc.org