Databases and Algorithms for Pathway Bioinformatics

Peter D. Karp, Ph.D.

Bioinformatics Research Group

SRI International

pkarp@ai.sri.com

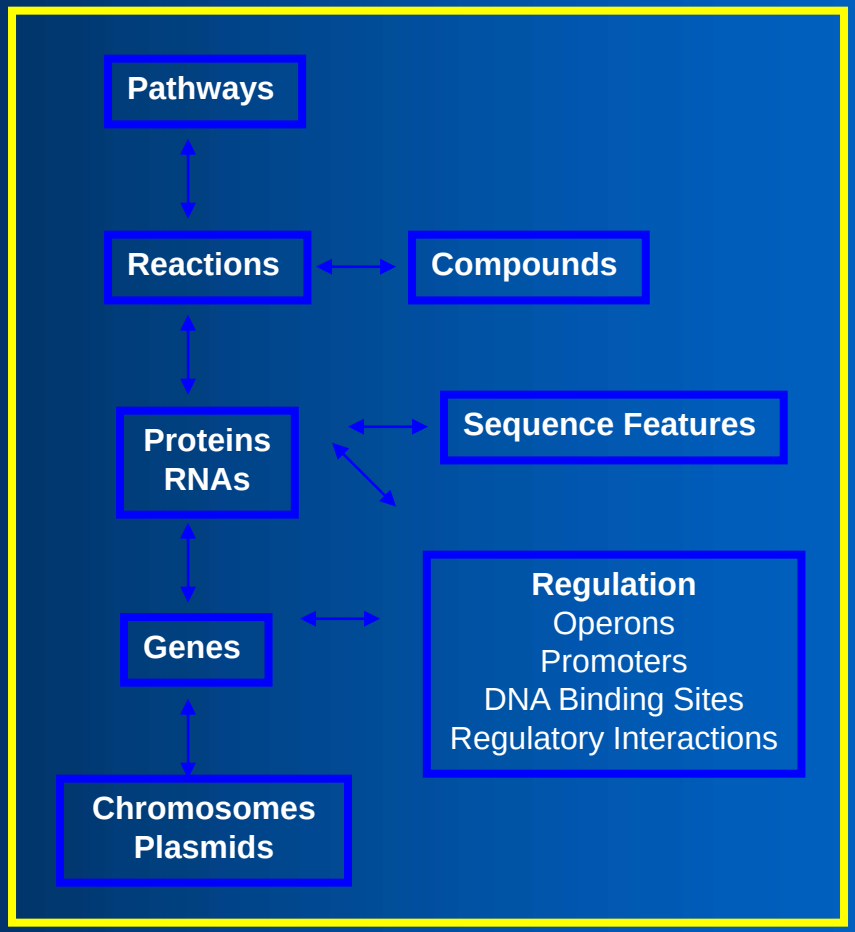
BioCyc.org

EcoCyc.org, MetaCyc.org, HumanCyc.org

Motivations: Management of Metabolic Pathway Data

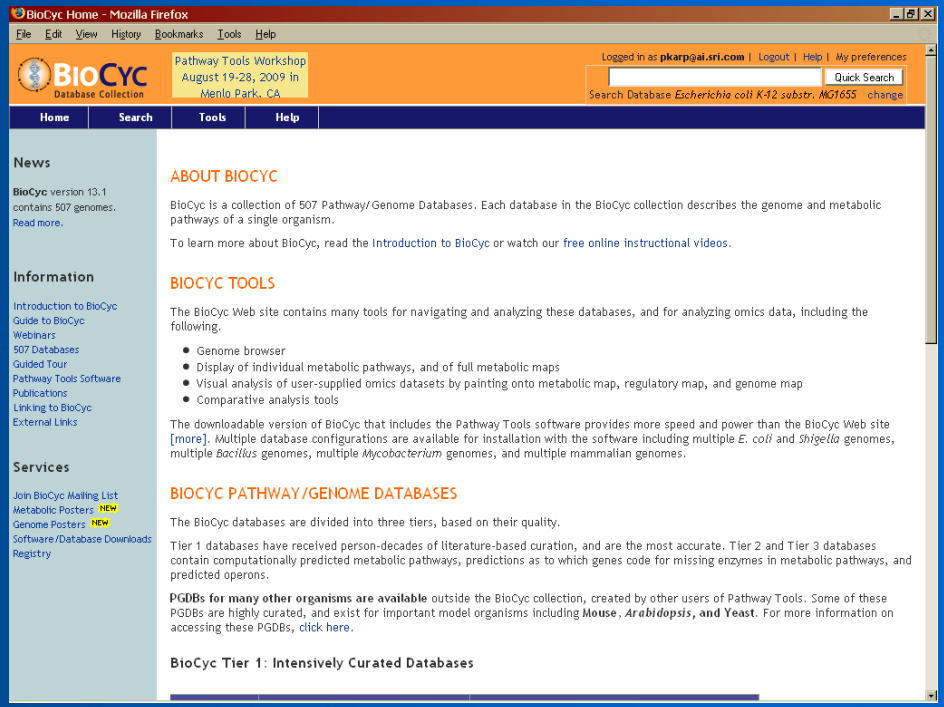
- **Organize growing corpus of data on metabolic pathways**
  - Experimentally elucidated pathways in the biomedical literature
  - Computationally predicted pathways derived from genome data
- **Provide software tools for querying and comprehending this complex information space**
- **Multiorganism view: MetaCyc**
  - Unique, experimentally elucidated pathways across all organisms
  - Reference database for computational pathway prediction
- **Organism-specific view:**
  - Organism-specific Pathway/Genome Databases
  - Detailed qualitative models of metabolic networks
  - Combine computational predictions with experimentally determined pathways

Pathway/Genome Database

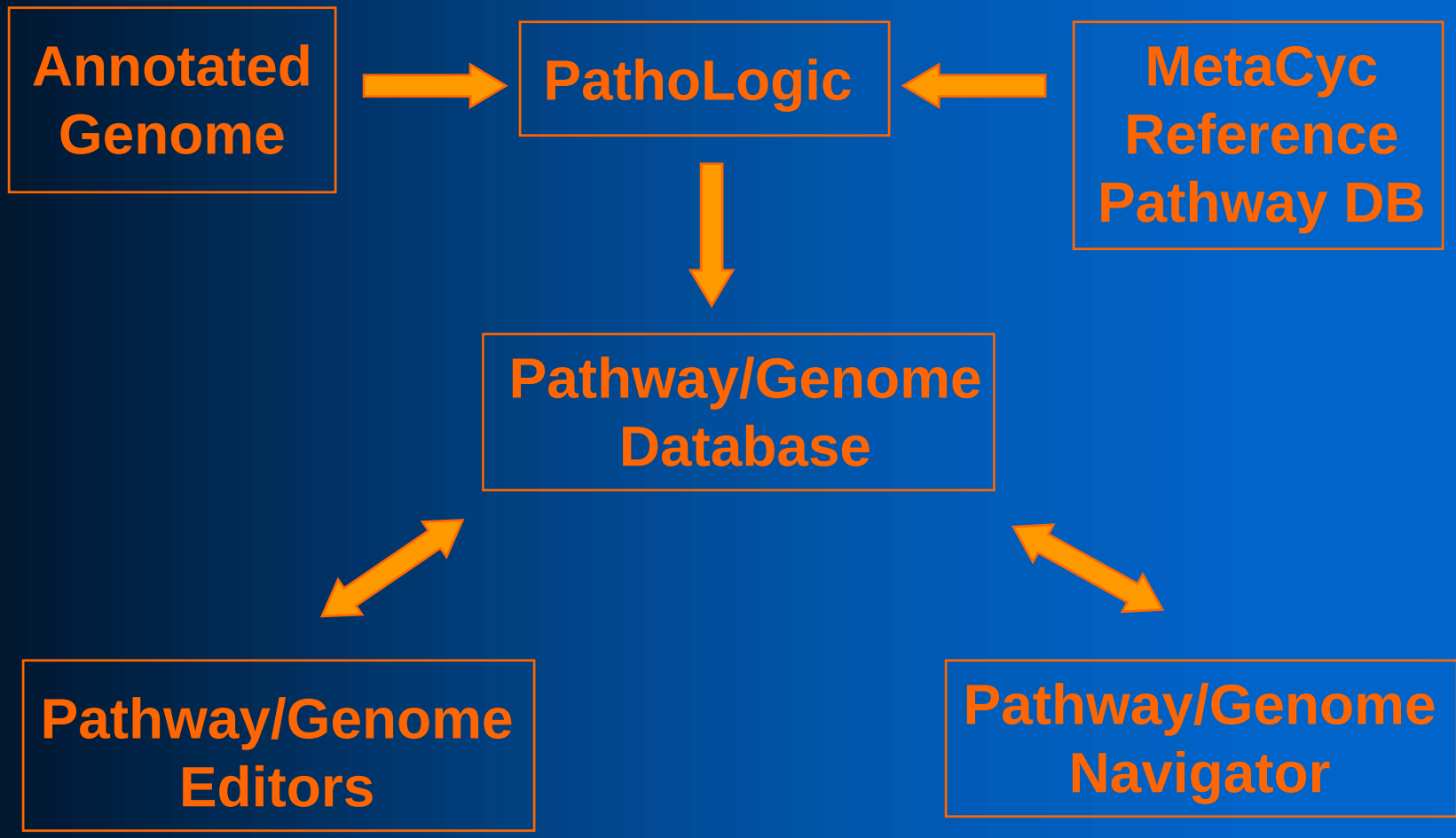


BioCyc Collection of 507 Pathway/Genome Databases

- **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 -- 24 PGDBs**
  - HumanCyc
  - Mycobacterium tuberculosis
- **Tier 3: Computationally-derived DBs, No Curation -- 481 DBs**



Pathway Tools Overview



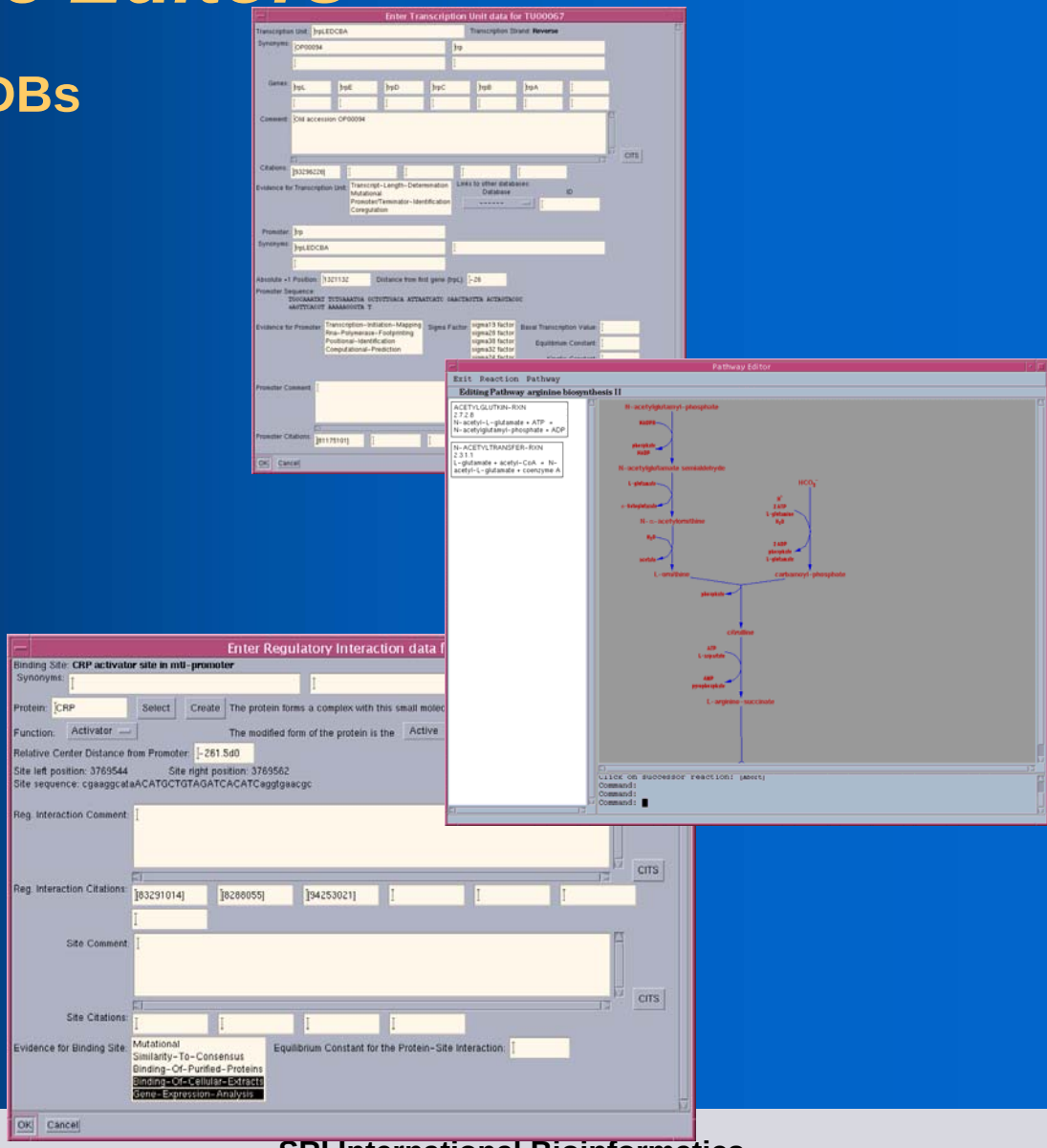
Pathway Tools Software: PathoLogic

- **Computational creation of new Pathway/Genome Databases**
- **Transforms genome into Pathway Tools schema and layers inferred information above the genome**
- **Predicts operons**
- **Predicts metabolic network**
- **Predicts which genes code for missing enzymes in metabolic pathways**
- **Infers transport reactions from transporter names**

Karp et al, *Briefings in Bioinformatics* 2009

Pathway Tools Software: Pathway/Genome Editors

- Interactively update PGDBs with graphical editors
- Support geographically distributed teams of curators with object database system
- Gene editor
- Protein editor
- Reaction editor
- Compound editor
- Pathway editor
- Operon editor
- Publication editor

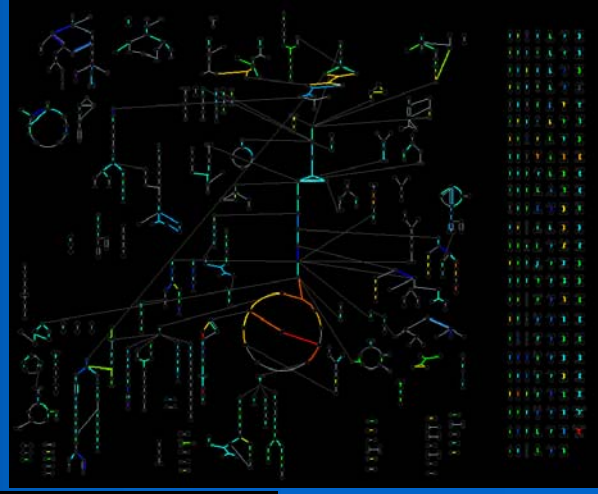


BioCyc Database Collection

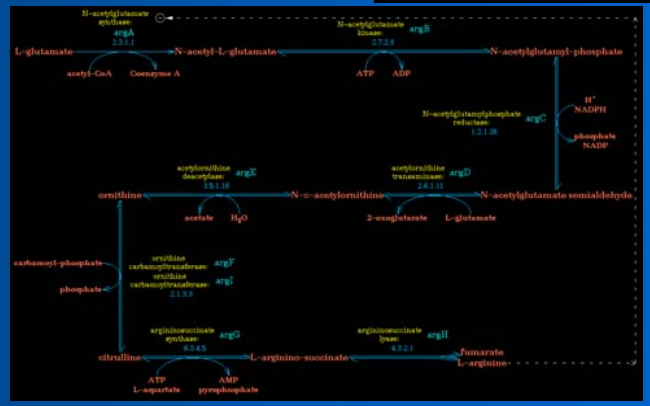
What is Curation?

- Ongoing updating and refinement of a PGDB
- Correcting false-positive and false-negative predictions
- Incorporating information from experimental literature
- Authoring of comments and citations
- Updating database fields
- Gene positions, names, synonyms
- Protein functions, activators, inhibitors
- Addition of new pathways, modification of existing pathways
- Defining TF binding sites, promoters, regulation of transcription initiation and other processes

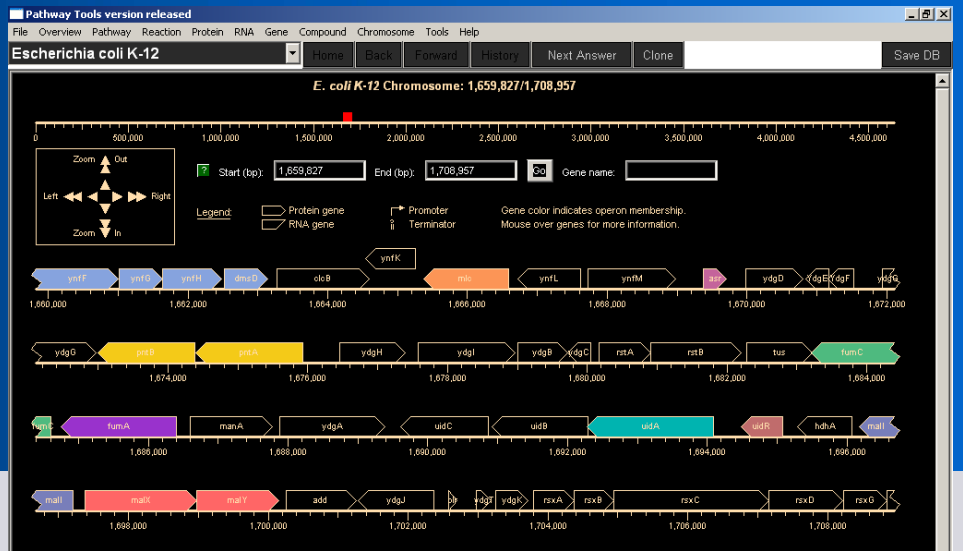
Pathway Tools Software: Pathway/Genome Navigator



- **Querying and visualization of:**
  - Pathways
  - Reactions
  - Metabolites
  - Proteins
  - Genes
  - Chromosomes



- **Two modes of operation:**
  - Web mode
  - Desktop mode
  - Most functionality shared, but each has unique functionality



BioCyc Database Collection

Pathway Tools Software: PGDBs Created Outside SRI

- **2,000+ licensees: 75+ groups applying software to 300+ organisms**

- ***Saccharomyces cerevisiae*, SGD project, Stanford University**
  - *135 pathways / 565 publications*
- ***Candida albicans*, CGD project, Stanford University**
- **dictyBase, Northwestern University**

- **Mouse, MGD, Jackson Laboratory**
- **Under development:**
  - *Drosophila,* FlyBase
  - *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**
- **Pathways, reactions, enzymes, substrates**

- **Jointly developed by**
  - P. Karp, R. Caspi, C. Fulcher, SRI International
  - L. Mueller, A. Pujar, Boyce Thompson Institute
  - S. Rhee, P. Zhang, Carnegie Institution

***Nucleic Acids Research* 2010**

MetaCyc Data -- Version 13.6

Pathways	1,436
Reactions	8,200
Enzymes	6,060
Small Molecules	8,400
Organisms	1,800
Citations	21,700

Taxonomic Distribution of MetaCyc Pathways – version 13.1

Bacteria	883
Green Plants	607
Fungi	199
Mammals	159
Archaea	112

- **Biosynthesis [902]**
  - Amino acids Biosynthesis [105]
  - Aromatic Compounds Biosynthesis [13]
  - Carbohydrates Biosynthesis [70]
  - Cell structures Biosynthesis [31]
  - Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis [160]
  - Hormones Biosynthesis [40]
  - Fatty Acids and Lipids Biosynthesis [101]
  - Metabolic Regulators Biosynthesis [4]
  - Nucleosides and Nucleotides Biosynthesis [20]
  - Amines and Polyamines Biosynthesis [32]
  - Secondary Metabolites Biosynthesis [351]
    - Antibiotic Biosynthesis [20]
    - Fatty Acid Derivatives Biosynthesis [7]
    - Flavonoids Biosynthesis [70]
    - Nitrogen-Containing Secondary Compounds Biosynthesis [64]
      - Alkaloids Biosynthesis [43]
    - Phenylpropanoid Derivatives Biosynthesis [46]
    - Phytoalexins Biosynthesis [25]
    - Sugar Derivatives Biosynthesis [10]
    - Terpenoids Biosynthesis [103]
  - Siderophore Biosynthesis [7]

- **Degradation/Utilization/Assimilation [639]**
  - Alcohols Degradation [14]
  - Aldehyde Degradation [12]
  - Amines and Polyamines Degradation [40]
  - Amino Acids Degradation [113]
  - Aromatic Compounds Degradation [152]
  - C1 Compounds Utilization and Assimilation [24]
  - Carbohydrates Degradation [52]
  - Carboxylates Degradation [30]
  - Chlorinated Compounds Degradation [39]
  - Cofactors, Prosthetic Groups, Electron Carriers Degradation [2]
  - Fatty Acid and Lipids Degradation [18]
  - Inorganic Nutrients Metabolism [72]
    - Nitrogen Compounds Metabolism [15]
    - Phosphorus Compounds Metabolism [3]
    - Sulfur Compounds Metabolism [54]
  - Nucleosides and Nucleotides Degradation and Recycling [9]
  - Secondary Metabolites Degradation [58]
    - Nitrogen Containing Secondary Compounds Degradation [13]
    - Sugar Derivatives Degradation [31]
    - Terpenoids Degradation [10]

- **Detoxification [16]**
  - Acid Resistance [2]
  - Arsenate Detoxification [3]
  - Mercury Detoxification [1]
  - Methylglyoxal Detoxification [8]

- **Generation of precursor metabolites and energy [124]**
  - Chemoautotrophic Energy Metabolism [14]
    - Hydrogen Oxidation [2]
  - Electron Transfer [11]
  - Fermentation [34]
  - Glycolysis [6]
  - Methanogenesis [12]
  - Pentose Phosphate Pathways [4]
  - Photosynthesis [6]
  - Respiration [25]
    - Aerobic Respiration [9]
    - Anaerobic Respiration [14]
  - TCA cycle [9]

What is a Pathway?

- A connected sequence of biochemical reactions
- Occurs in one organism
- Conserved through evolution
- Regulated as a unit
- Often starts or stops at one of 13 common intermediate metabolites

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**

Enzyme Data Available in MetaCyc

- Reaction(s) catalyzed
- Alternative substrates
- Activators, inhibitors, cofactors, prosthetic groups
- Subunit structure
- Genes
- Features on protein sequence
- Cellular location
- pI, molecular weight, Km, Vmax
- Gene Ontology terms
- Links to other bioinformatics databases

Comparison with KEGG

- **KEGG vs MetaCyc: Reference pathway collections**
  - KEGG maps are not pathways *Nuc Acids Res* 34:3687 2006
    - KEGG maps contain multiple biological pathways
    - Two genes chosen at random from a BioCyc pathway are more likely to be related according to genome context methods than from a KEGG pathway
    - KEGG 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 assigns half as many reactions to pathways as MetaCyc

- **KEGG vs organism-specific PGDBs**
  - KEGG does not curate or customize pathway networks for each organism
  - Highly curated PGDBs now exist for important organisms such as E. coli, yeast, mouse, Arabidopsis

PathoLogic Step 3: Prediction of Metabolic Pathways

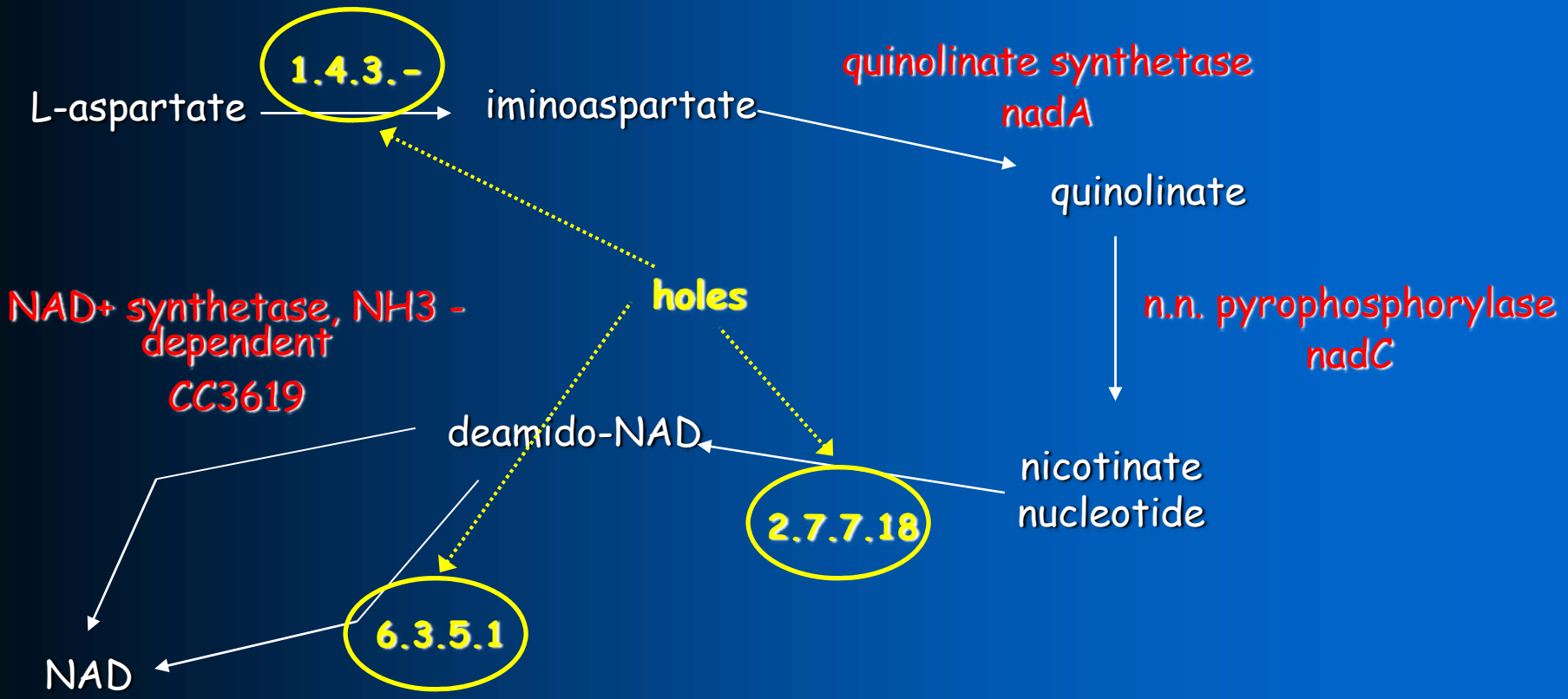
- **Infer reaction complement of organism**
  - Match enzymes in source genome to MetaCyc reactions by enzyme name, EC number, GO term
  - Support user in manually matching additional enzymes

- **Computationally predict which MetaCyc metabolic pathways are present**
  - For each MetaCyc pathway, evaluate which of its reactions are catalyzed by the organism
  - Features: Fraction of reactions present, number of unique reactions, taxonomic domain of pathway
  - Many other features explored with machine learning methods

BMC Bioinformatics 2009

PathoLogic Step 4: Pathway Hole Filler

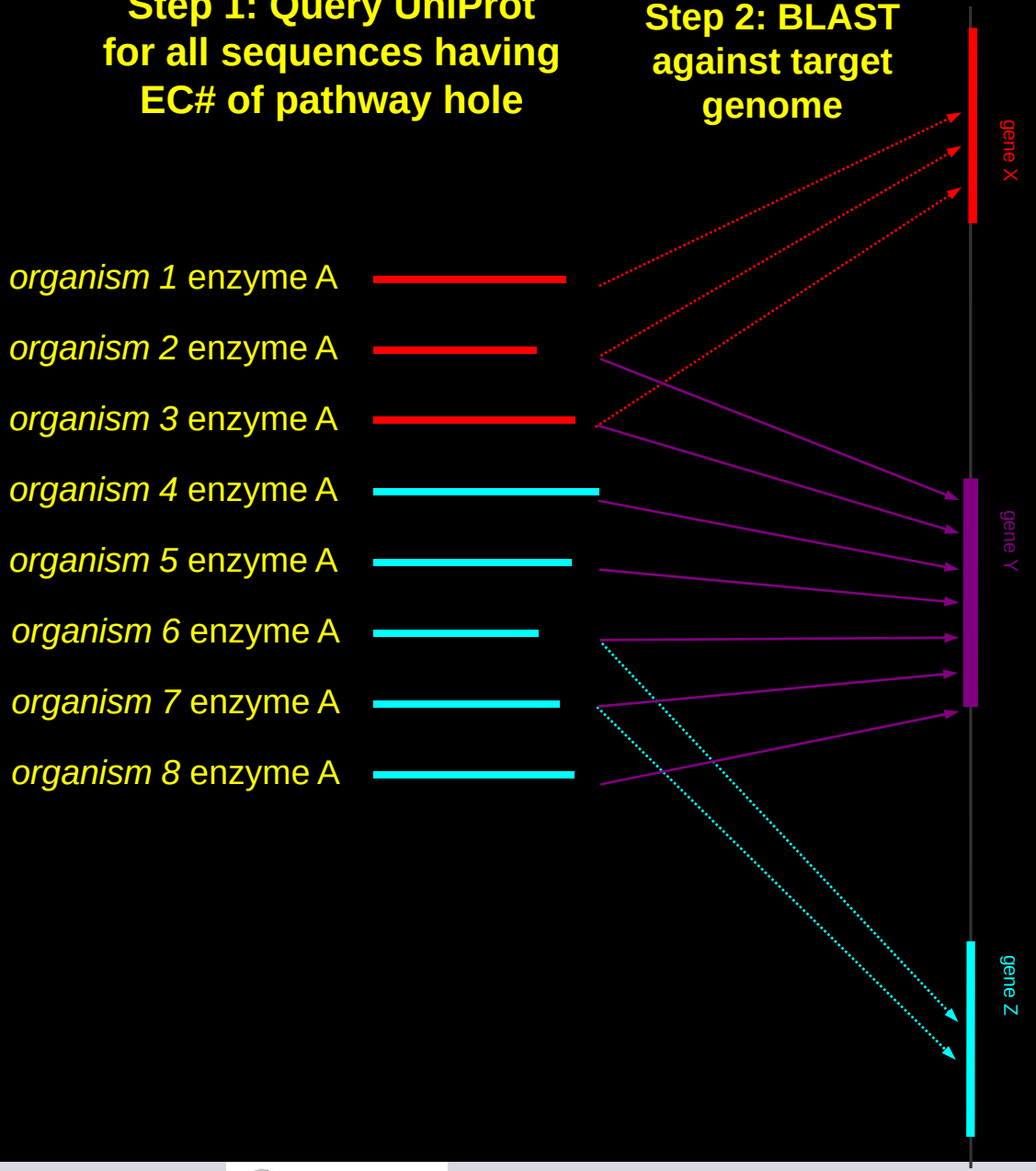
- **Definition: Pathway Holes are reactions in metabolic pathways for which no enzyme is identified**



BioCyc Database Collection

Step 1: Query UniProt for all sequences having EC# of pathway hole

Step 2: BLAST against target genome

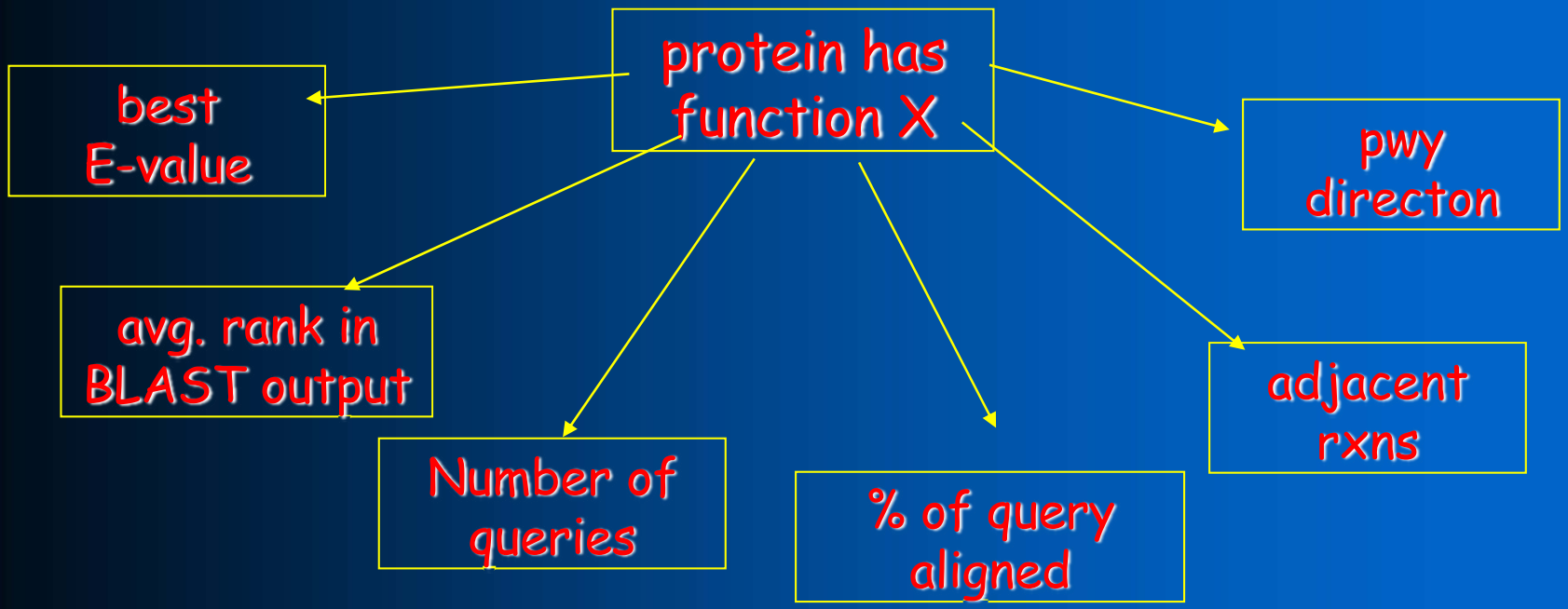


Step 3 & 4: Consolidate hits and evaluate evidence

7 queries have high-scoring hits to sequence Y

Bayes Classifier

P(protein has function X|
E-value, avg. rank, aln. length, etc.)



BMC Bioinformatics 5:76 2004

PathoLogic Step 5: Transport Inference Parser

- **Problem: Write a program to query a genome annotation to compute the substrates an organism can transport**

- **Typical genome annotations for transporters:**
  - ATP transporter for ribose
  - ribose ABC transporter
  - D-ribose ATP transporter
  - ABC transporter, membrane spanning protein [ribose]
  - ABC transporter, membrane spanning protein [D-ribose]

Transport Inference Parser

- **Input: "ATP transporter of phosphonate"**
- **Output: Structured description of transport activity**

- **Locates most transporters in genome annotation using keyword analysis**

- **Parse product name using a series of rules to identify:**
  - Transported substrate, co-substrate
  - Influx/efflux
  - Energy coupling mechanism

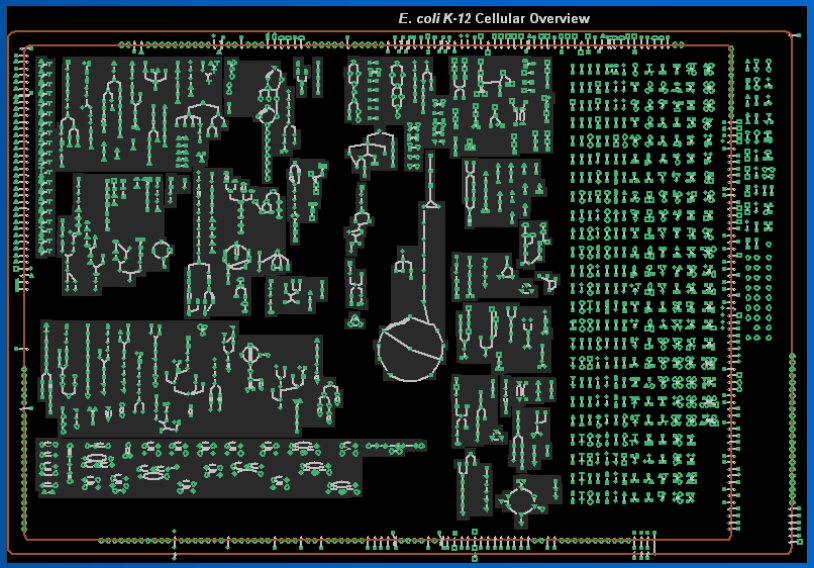
- **Creates transport reaction object:**

$phosphonate_{[periplasm]} + H_2O + ATP = phosphonate + P_i + ADP$

Tools

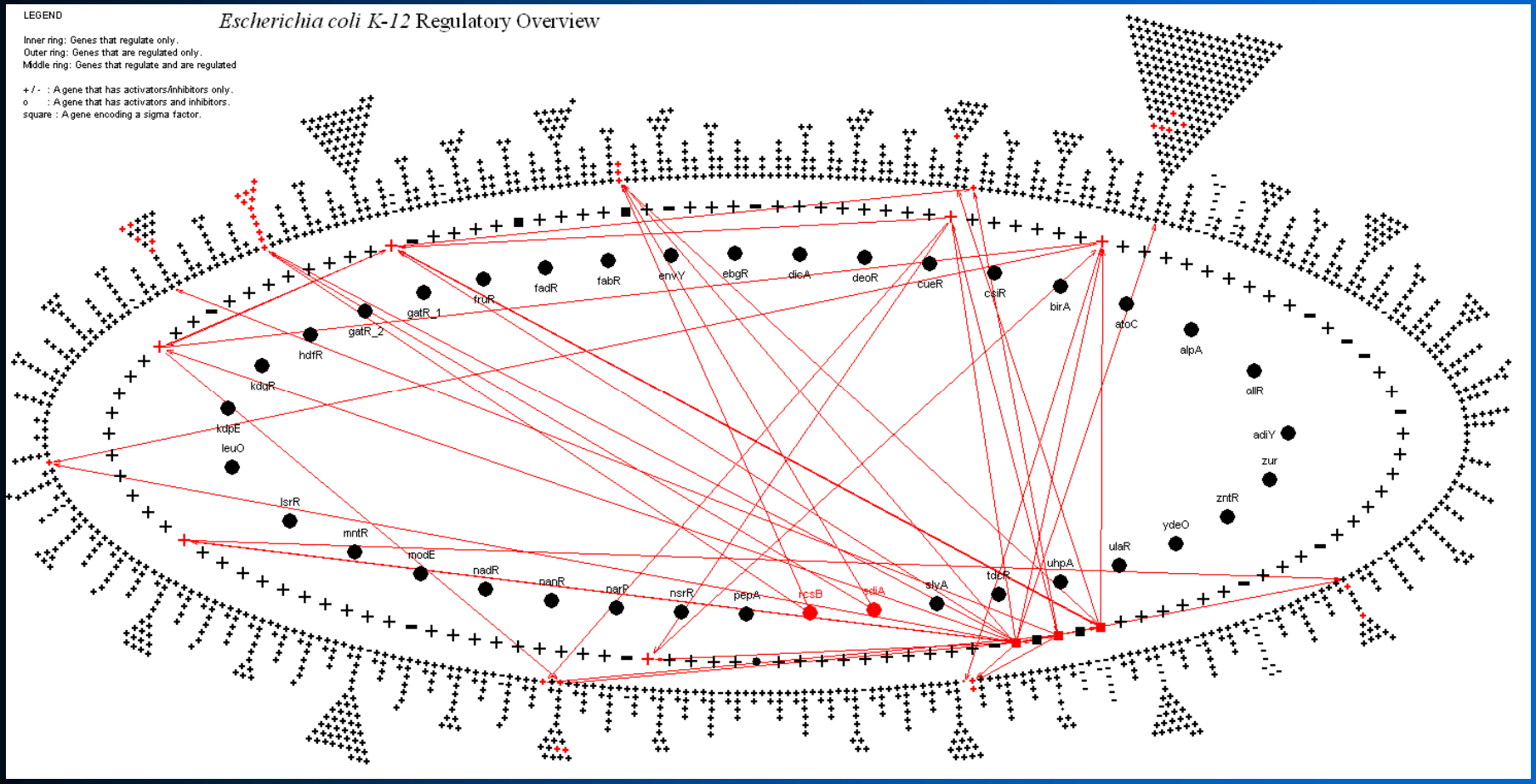
Pathway Tools Overviews and Omics Viewers

- **Genome-scale visualizations of cellular networks**
- **Harness human visual system to interpret patterns in biological contexts**
- **Designed to avoid the hairball effect**
- **Generated automatically from PGDB**
- **Magnify, interrogate**
- **Omics viewers paint omics data onto overview diagrams**
  - Different perspectives on same dataset
  - Use animation for multiple time points or conditions
  - Paint any data that associates numbers with genes, proteins, reactions, or metabolites



Regulatory Overview and Omics Viewer

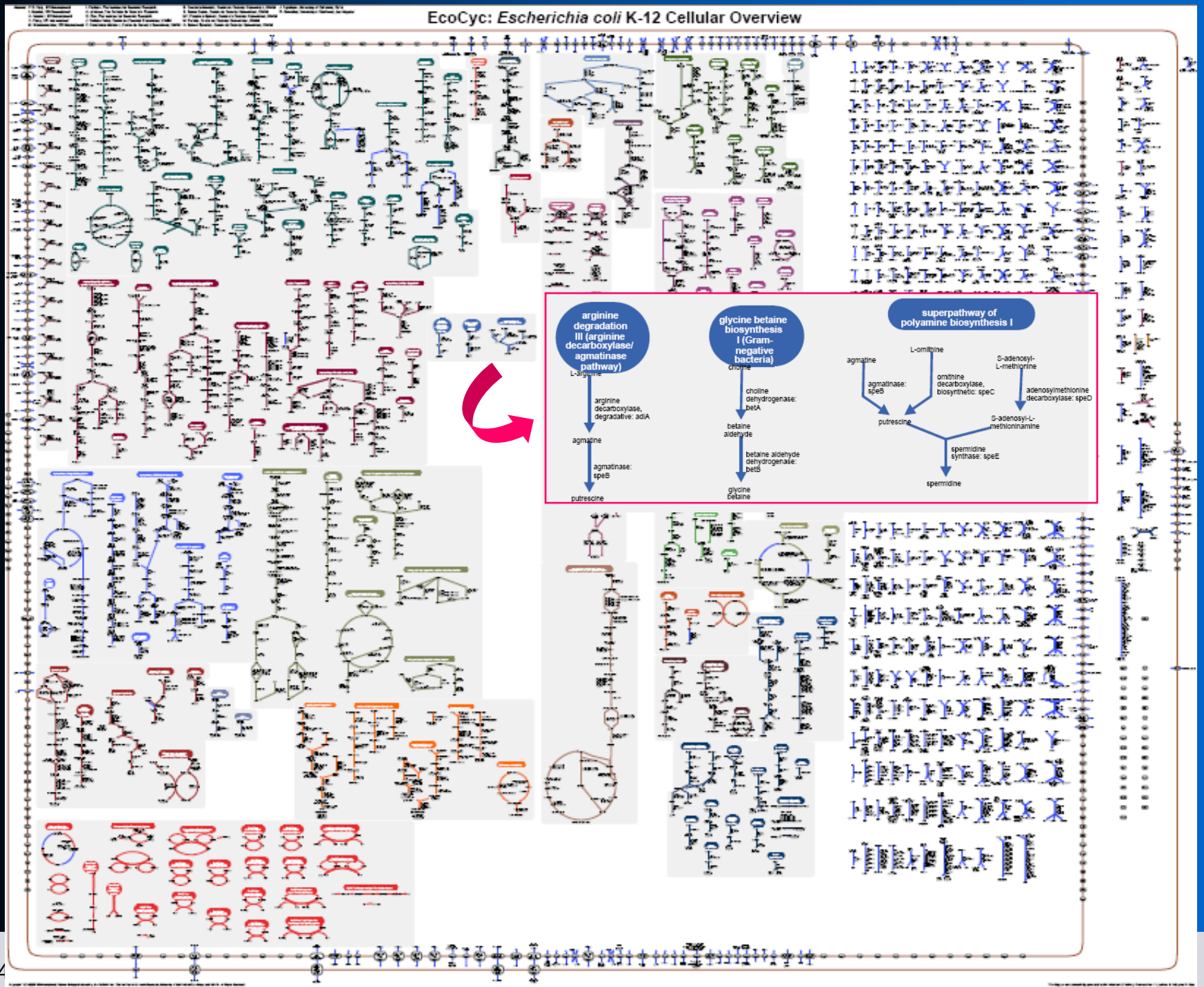
- Show regulatory relationships among gene groups



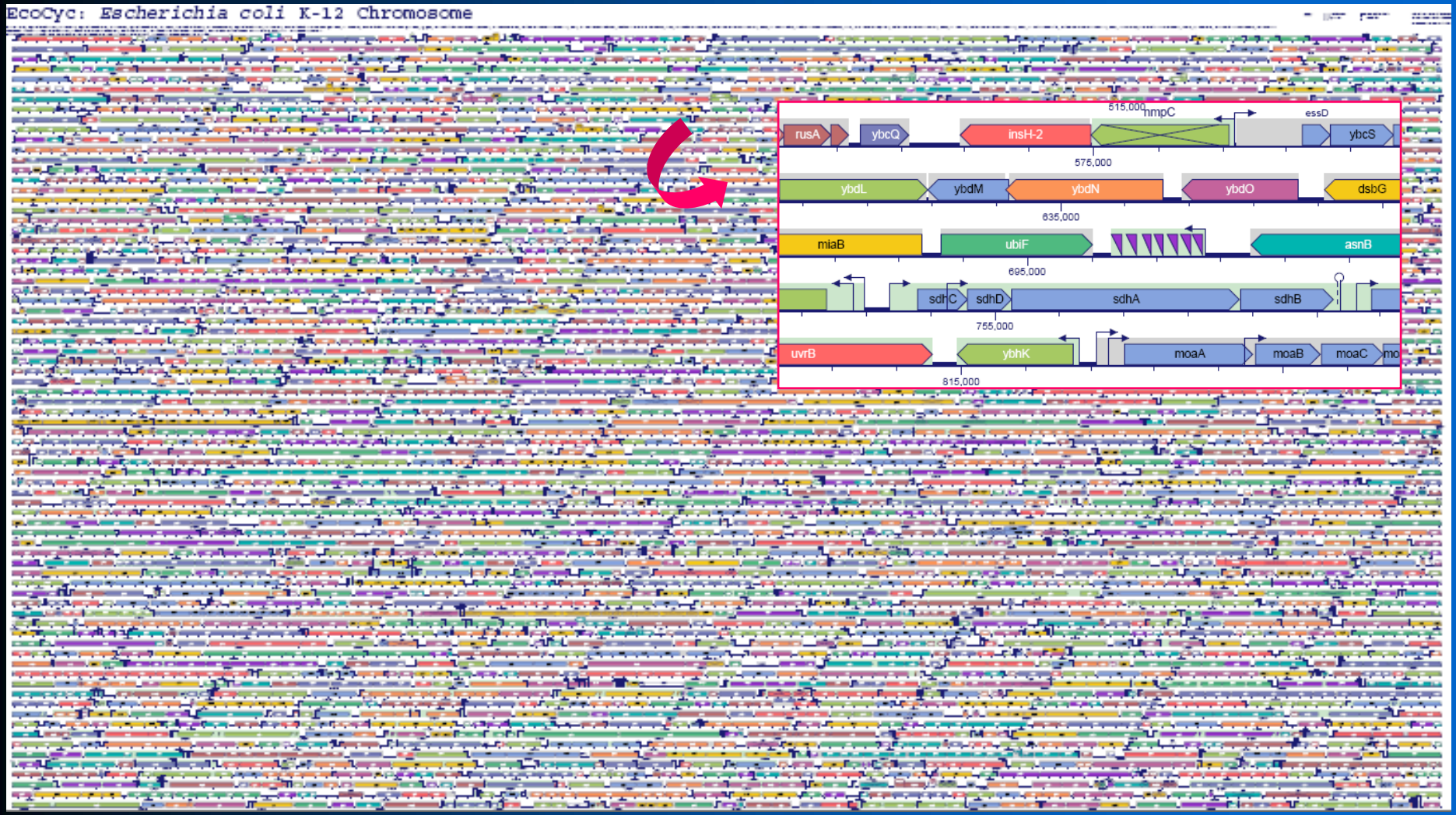
BioCyc Database Collection

SRI International Bioinformatics

Genome Overview


EcoCyc: Escherichia coli K-12 Cellular Overview
arginine degradation III (arginine decarboxylase/ agmatinase pathway)
L-arginine
arginine decarboxylase, degradative: adiA
agmatine
agmatinase: speB
putrescine
glycine betaine biosynthesis I (Gram-negative bacteria)
choline
choline dehydrogenase: betA
betaine aldehyde
betaine aldehyde dehydrogenase: betB
glycine betaine
superpathway of polyamine biosynthesis I
L-ornithine
agmatine
S-adenosyl-L-methionine
agmatinase: speB
ornithine decarboxylase, biosynthetic: speC
adenosylmethionine decarboxylase: speD
putrescine
S-adenosyl-L-methioninamine
spermidine synthase: speE
spermidine

Genome Poster



BioCyc Database Collection

Dead End Metabolite Finder

- **A small molecule C is a dead-end if:**
  - C is produced only by SMM reactions in Compartment, and no transporter acts on C in Compartment OR
  - C is consumed only by SMM reactions in Compartment, and no transporter acts on C in Compartment

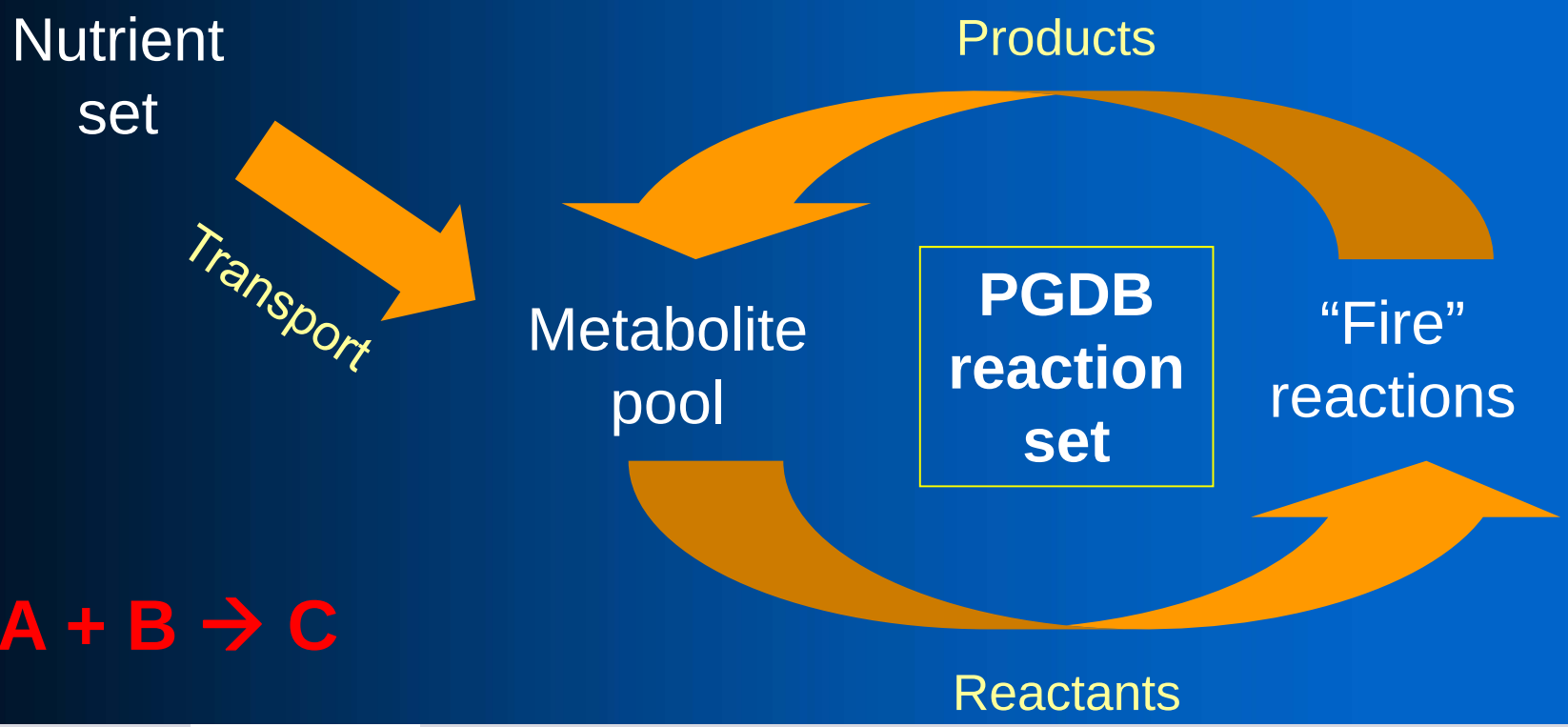
Reachability Analysis of Metabolic Networks

- **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**



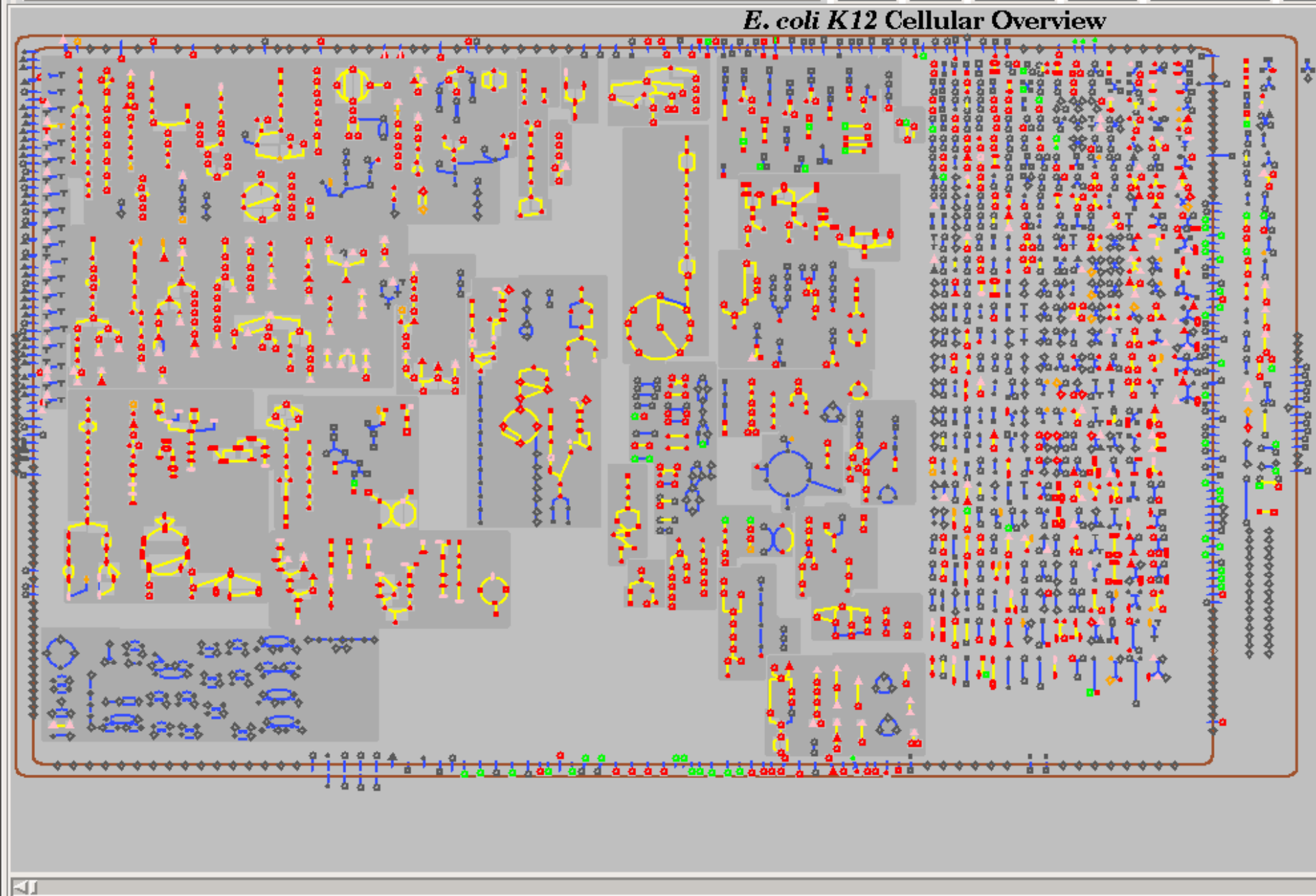
A + B → C

Initial Metabolite Nutrient Set (Total: 21 compounds)

Nutrients (8) (M61 Minimal growth medium)	H^+, Fe^{2+}, Mg^{2+}, K^+, NH_3, SO_4^{2-}, PO_4^{2-}, Glucose
Nutrients (10) (Environment)	Water, Oxygen, Trace elements (Mn^{2+}, Co^{2+}, Mo^{2+}, Ca^{2+}, Zn^{2+}, Cd^{2+}, Ni^{2+}, Cu^{2+})
Bootstrap Compounds (3)	ATP, NADP, CoA

Essential Compounds
E. coli Total: 41 compounds

- **Proteins (20)**
  - Amino acids
- **Nucleic acids (DNA & RNA) (8)**
  - Nucleosides
- **Cell membrane (3)**
  - Phospholipids
- **Cell wall (10)**
  - Peptidoglycan precursors
  - Outer cell wall precursors (Lipid-A, oligosaccharides)


E. coli K12 Cellular Overview

Results from EcoCyc Reachability Analysis in 2001

- **Phase I: Forward propagation**
  - 21 initial compounds yielded only half of the 41 essential compounds for *E. coli*
- **Phase II: Manually identify**
  - Bugs in EcoCyc (e.g., two objects for tryptophan)
    - A → B B' → C
  - Incomplete knowledge of *E. coli* metabolic network
    - A + B → C + D
  - "Bootstrap compounds"
  - Missing initial protein substrates (e.g., ACP)
    - Protein synthesis not represented
- **Phase III: Forward propagation with 11 more initial metabolites**
  - Yielded all 41 essential compounds

Regulation

Encoding Cellular Regulation in Pathway Tools -- Goals

- **Facilitate curation of wide range of regulatory information within a formal ontology**
- **Compute with regulatory mechanisms and pathways**
  - Summary statistics, complex queries
  - Pattern discovery
  - Visualization of network components
- **Provide training sets for inference of regulatory networks**
- **Interpret gene-expression datasets in the context of known regulatory mechanisms**

Regulatory Interactions Supported by Pathway Tools

- Substrate-level regulation of enzyme activity
- Binding to proteins or small molecules (phosphorylation)
- Regulation of transcription initiation
- Attenuation of transcription
- Regulation of translation by proteins and by small RNAs

Summary

- **Pathway/Genome Databases**
  - MetaCyc non-redundant DB of literature-derived pathways
  - 500 organism-specific PGDBs available through SRI at BioCyc.org
  - Additional curated PGDBs for mouse, yeast, Arabidopsis, etc
  - Computational theories of biochemical machinery

- **Pathway Tools software**
  - Predicts pathways and pathway hole fillers
  - Reachability analysis, dead-end metabolite analysis
  - Omics data analysis tools
  - Captures many bacterial regulatory interactions

BioCyc and Pathway Tools Availability

- **BioCyc.org Web site and database files freely available to all**
- **Pathway Tools freely available to non-profits**
  - Macintosh, PC/Windows, PC/Linux

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

BioCyc.org

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