

Pathway Bioinformatics

Peter D. Karp, PhD
Bioinformatics Research Group
SRI International
Menlo Park, CA

pkarp@ai.sri.com

BioCyc.org

Overview

- **Definitions**
- **BioCyc collection of Pathway/Genome Databases**
- **Algorithms for pathway bioinformatics**
- **Pathway Tools software**
 - Navigation and analysis
 - Infer metabolic pathways from genomes
- **Pathway Tools ontology**

Pathway Bioinformatics

- **The subfield of bioinformatics concerned with ontologies, algorithms, databases and visualizations of pathways**
- **Examples:**
 - Inference of metabolic pathways from genomes
 - Schemas for pathway DBs
 - Exchange formats for pathway data
 - Classification systems for pathway data
 - Pathway diagram layout algorithms

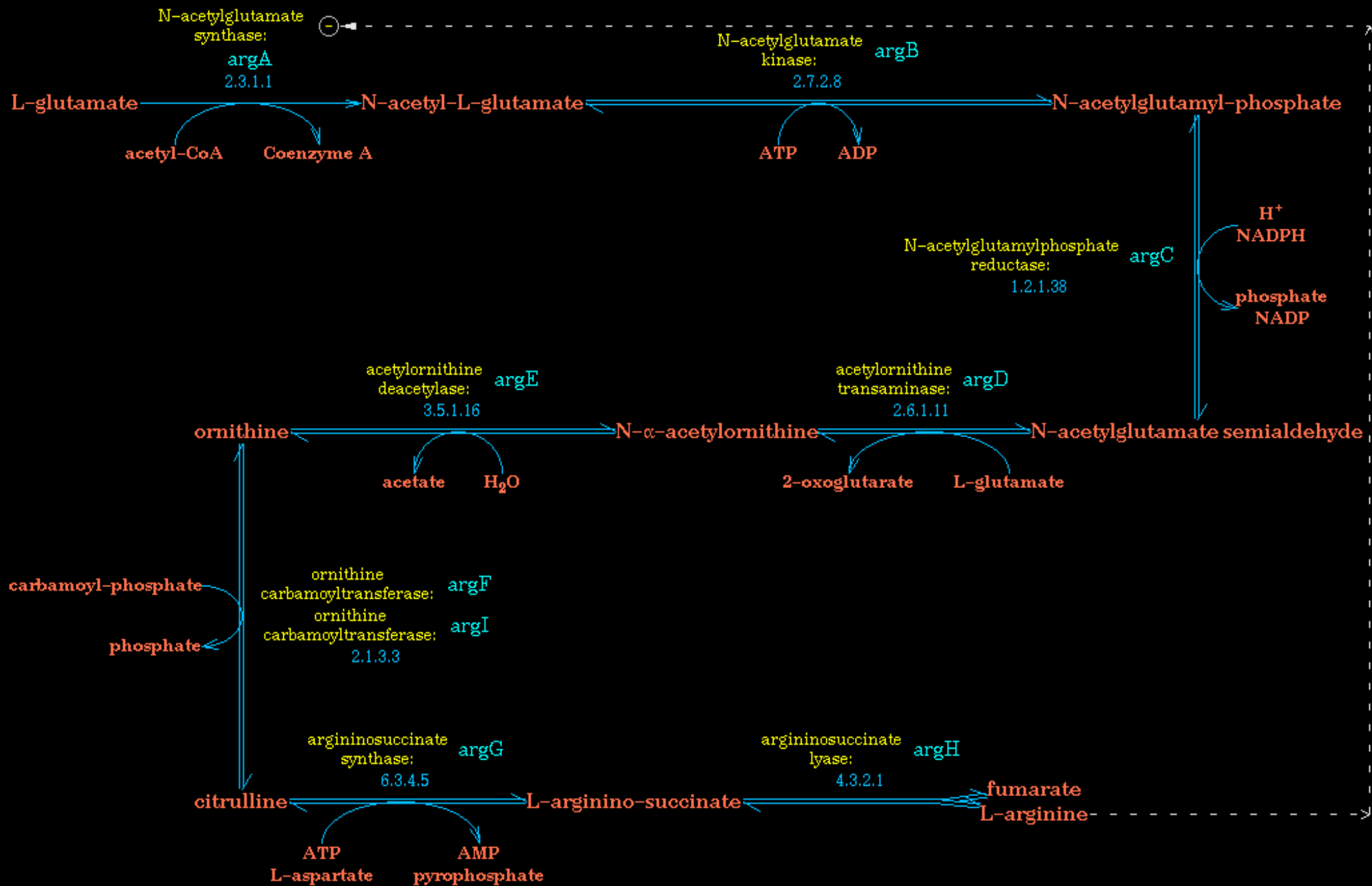
Definition of Metabolic Pathways

- A chemical reaction interconverts chemical compounds (analogous to a production rule)



- An enzyme is a protein that accelerates chemical reactions. Each enzyme is encoded by one or more genes.
- A pathway is a linked set of reactions (analogous to a chain of rules)

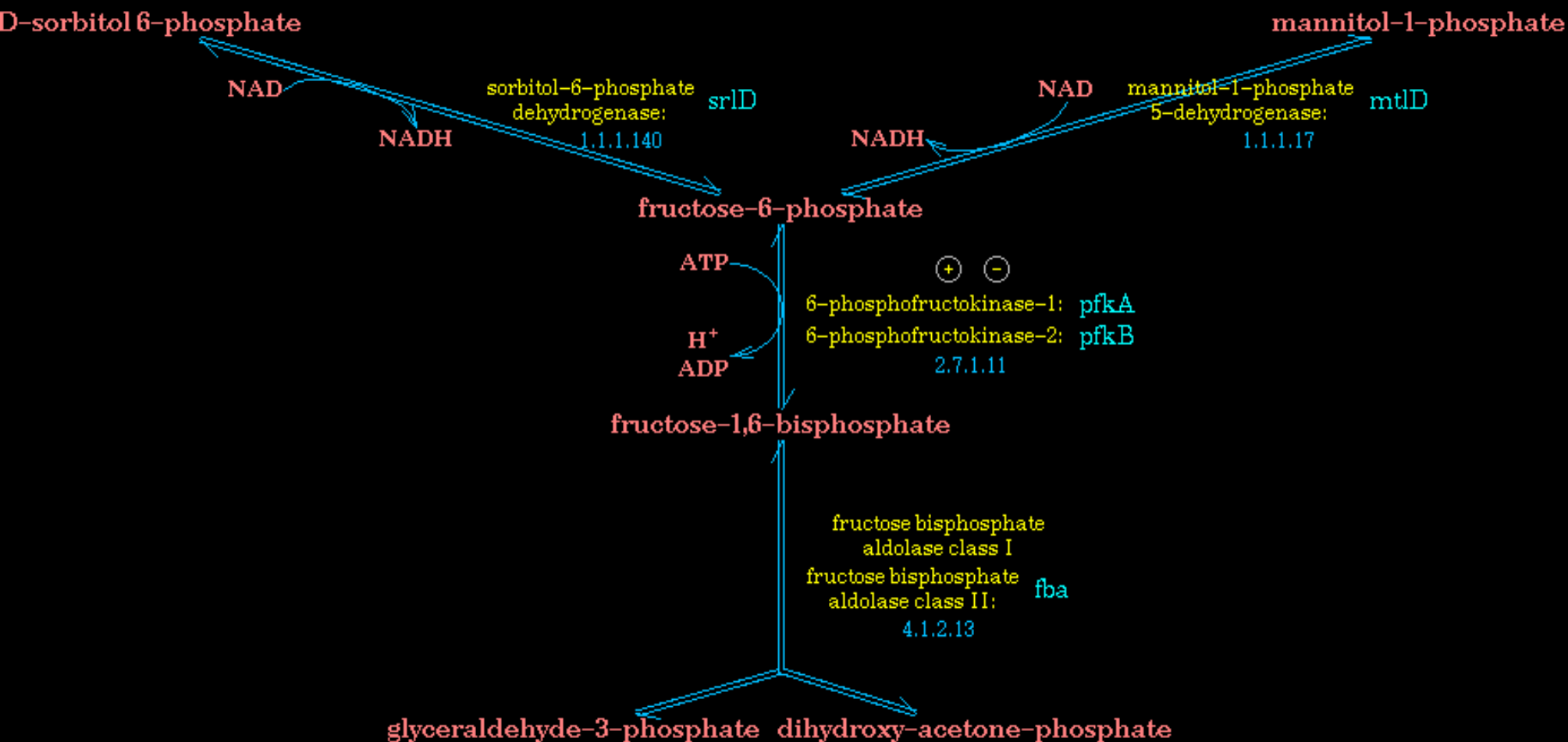




E. coli Pathway: hexitol degradation super-pathway

More Detail

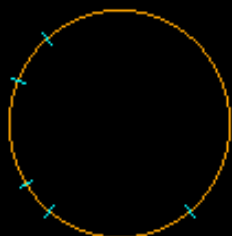
Less Detail



Superclasses: Carbon compounds, Super-Pathways

Subpathways: mannitol degradation, sorbitol degradation

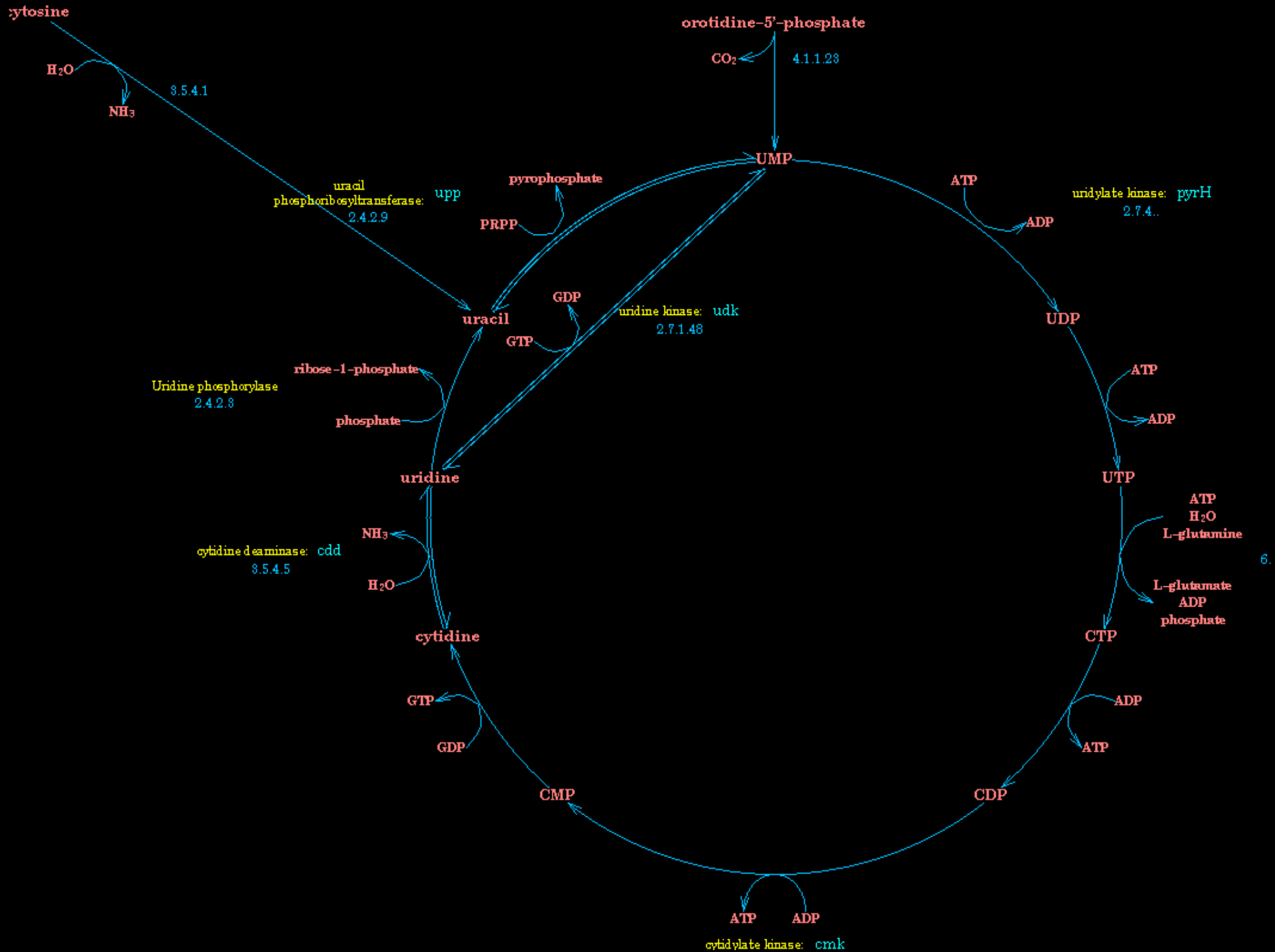
Locations of Mapped Genes:



Mp. pneumoniae Pathway: pyrimidine ribonucleotide/ribonucleoside metabolism

More Detail

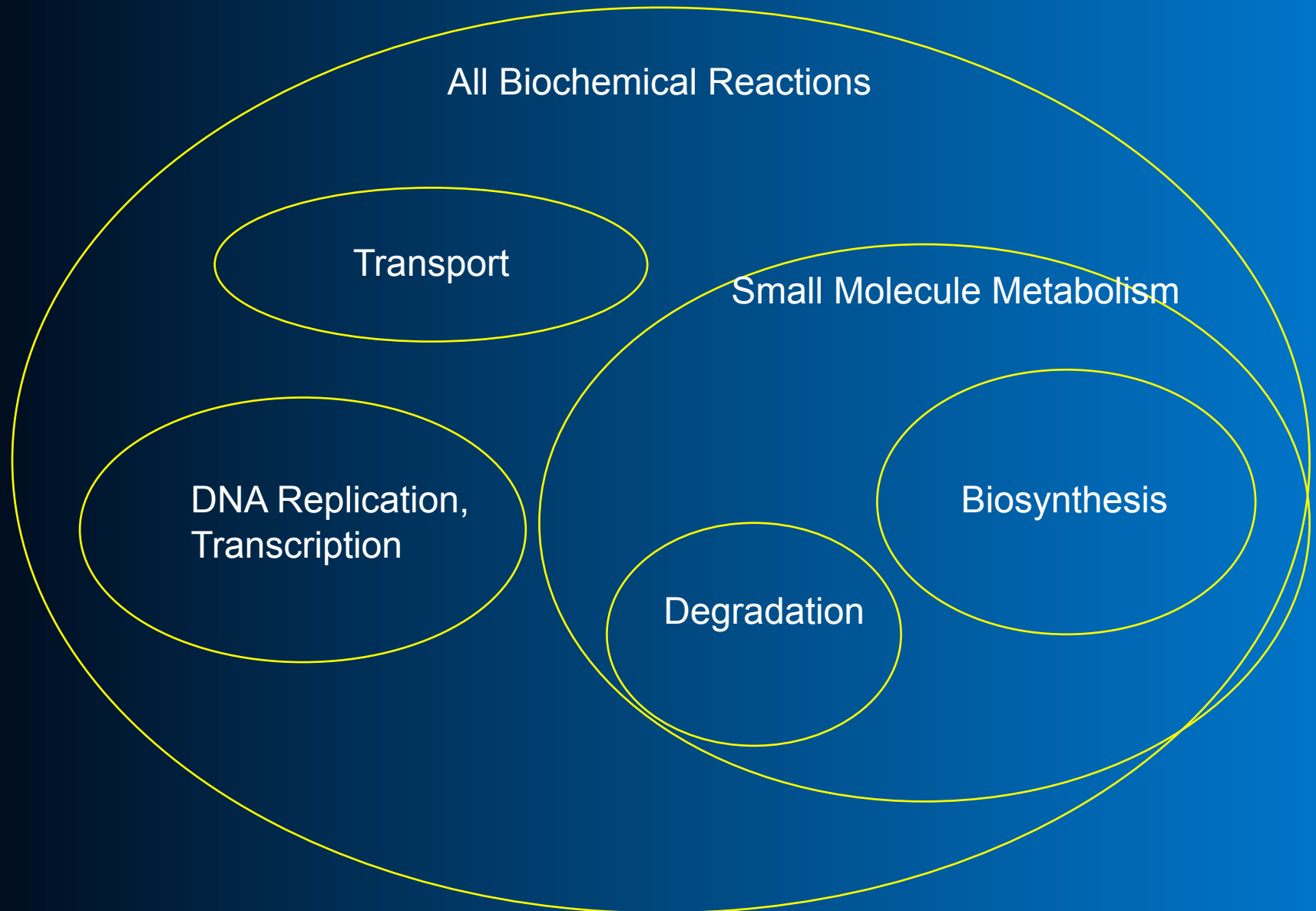
Less Detail



Definition of Small-Molecule Metabolism

- Small-molecule metabolism
 - Biochemical factory within the cell
 - Hundreds of enzyme-catalyzed reactions operating principally on small-molecule substrates

Small Molecule Metabolism

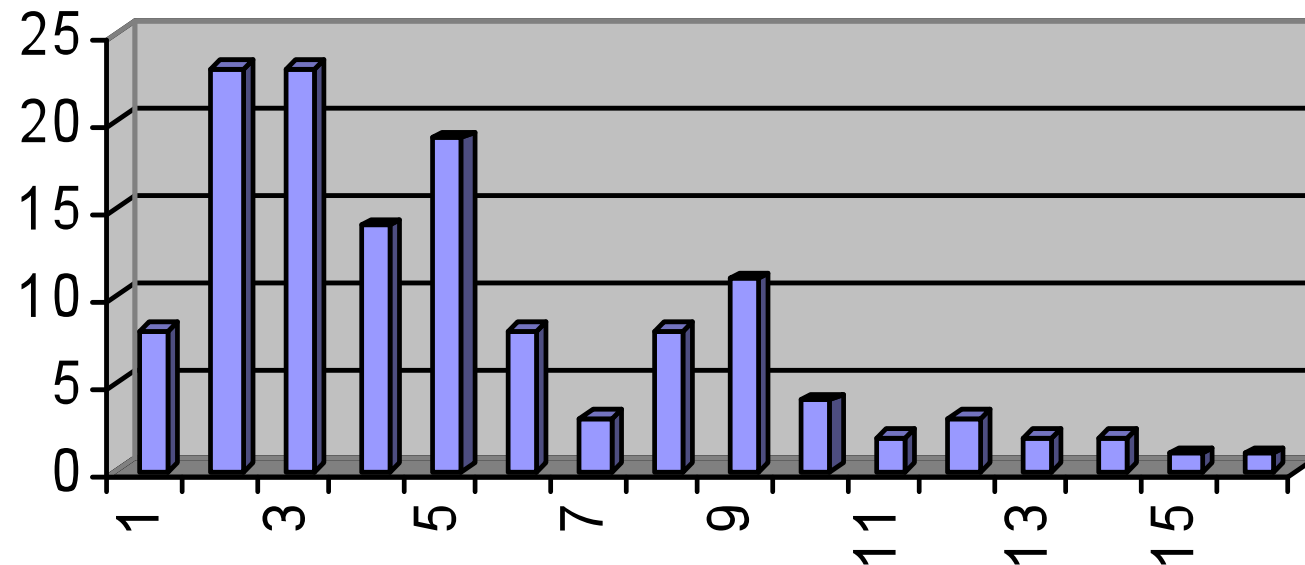


What is a Metabolic Pathway?

- **A pathway is a conceptual unit of the metabolism**
- **An ordered set of interconnected, directed biochemical reactions**
- **A pathway forms a coherent unit:**
 - Boundaries defined at high-connectivity substrates
 - Regulated as a single unit
 - Evolutionarily conserved across organisms as a single unit
 - Performs a single cellular function
 - Historically grouped together as a unit
 - All reactions in a single organism

EcoCyc Pathways

Pathway length distribution



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

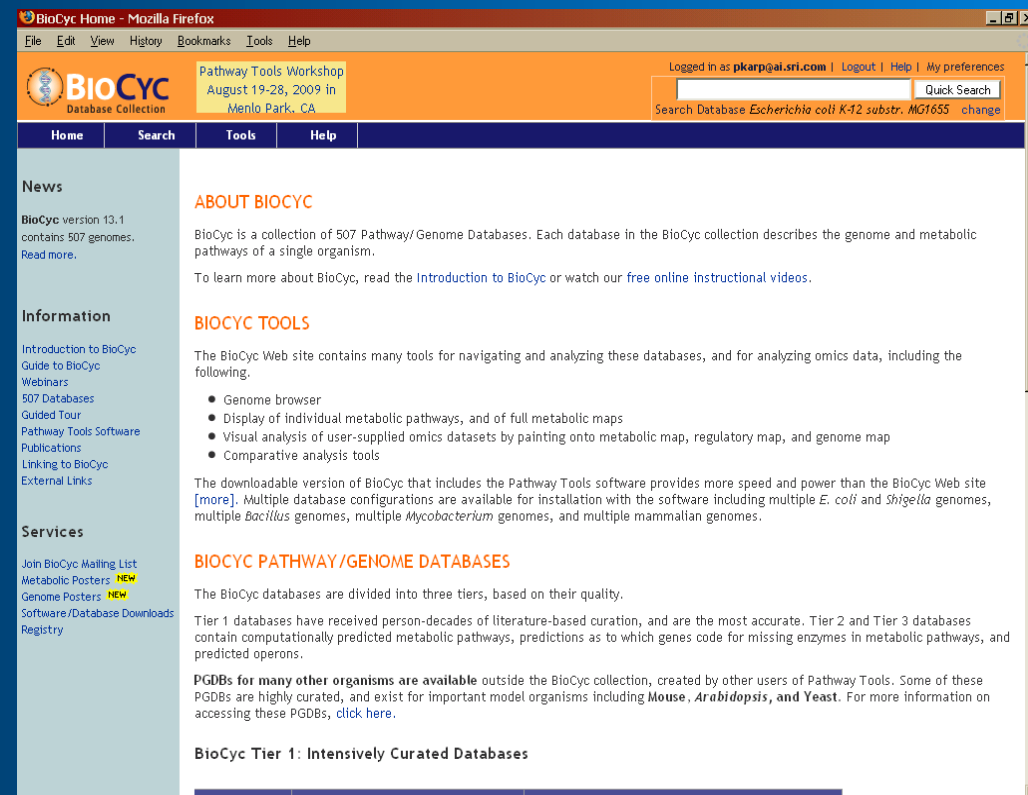
● Tier 1: Literature-Derived PGDBs

- 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



The screenshot shows the BioCyc website homepage. The browser title is "BioCyc Home - Mozilla Firefox". The page features a navigation menu with "Home", "Search", "Tools", and "Help". A sidebar on the left contains sections for "News", "Information", and "Services". The main content area includes "ABOUT BIOCYC", "BIOCYC TOOLS", "BIOCYC PATHWAY / GENOME DATABASES", and "BioCyc Tier 1: Intensively Curated Databases".

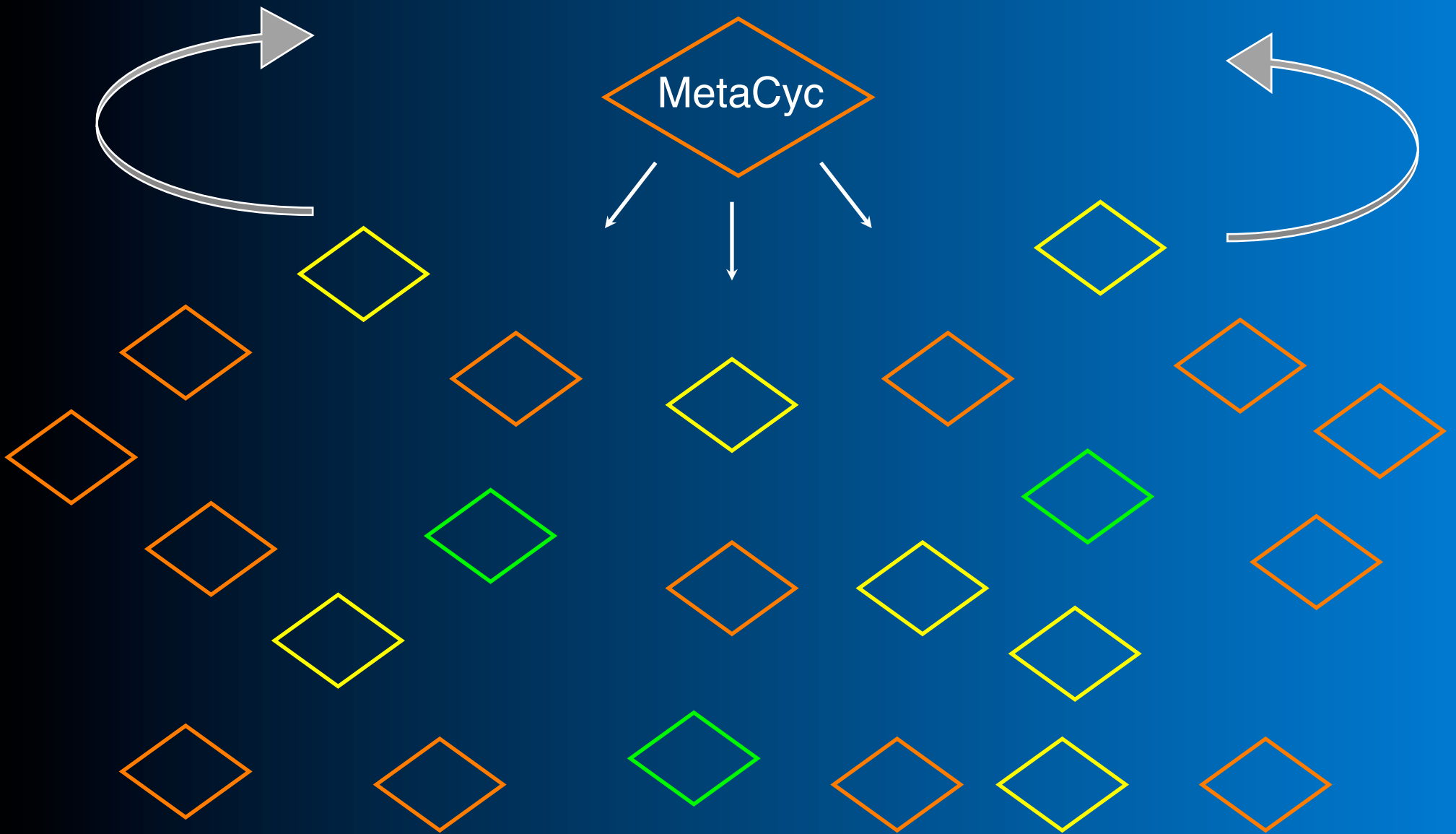
ABOUT BIOCYC
BioCyc is a collection of 507 Pathway/Genome Databases. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism.
To learn more about BioCyc, read the Introduction to BioCyc or watch our free online instructional videos.

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.

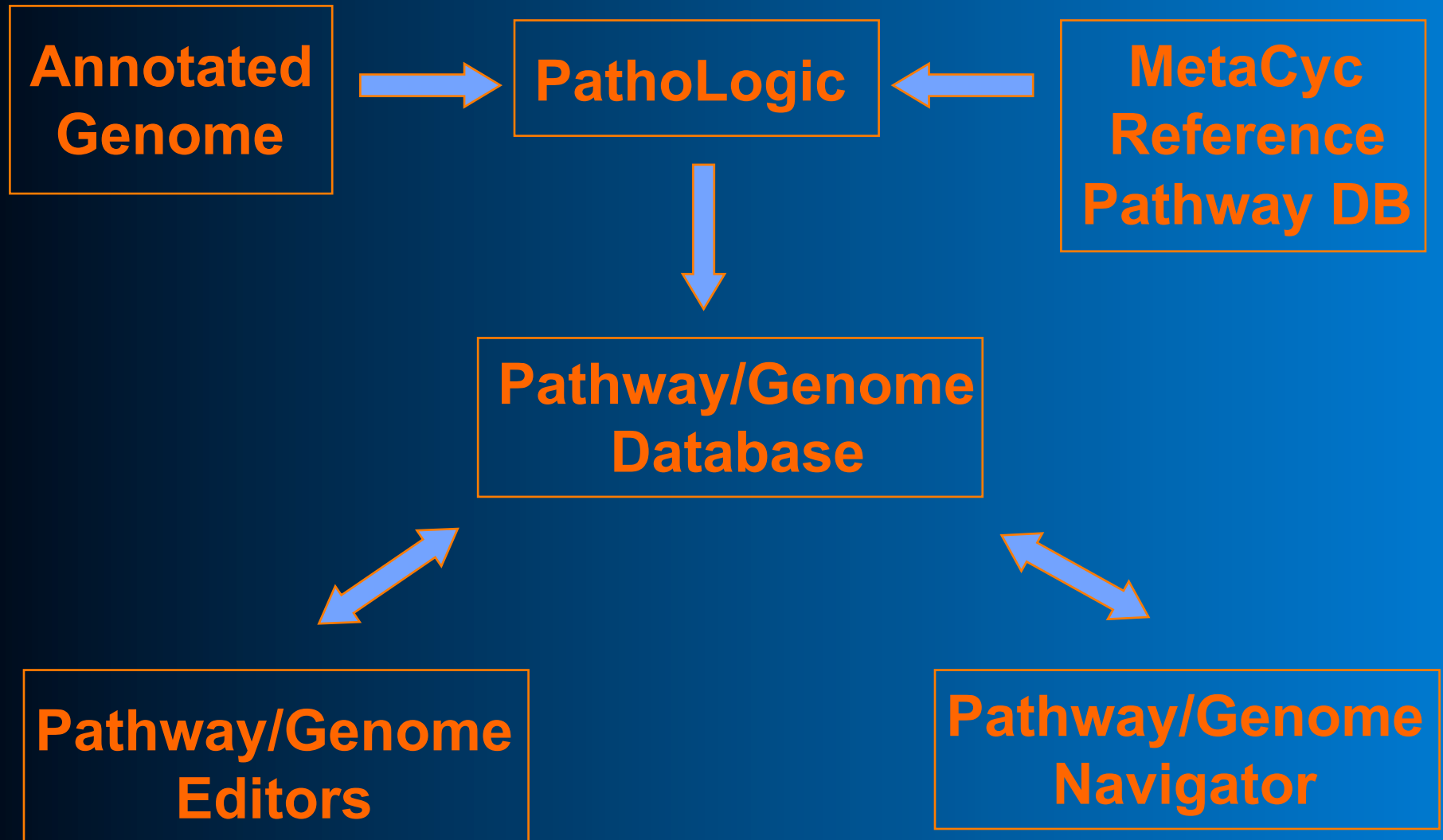
BIOCYC PATHWAY / GENOME DATABASES
The BioCyc databases are divided into three tiers, based on their quality.
Tier 1 databases have received person-decades of literature-based curation, and are the most accurate. Tier 2 and Tier 3 databases contain computationally predicted metabolic pathways, predictions 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

Family of Pathway/Genome Databases



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 pathway hole fillers**
- **Infers transport reactions**

MetaCyc: Metabolic Encyclopedia

- **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**
- **Jointly developed by SRI and Carnegie Institution**

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

MetaCyc Enzyme Data

- **Reaction(s) catalyzed**
- **Alternative substrates**
- **Cofactors / prosthetic groups**
- **Activators and inhibitors**
- **Subunit structure**
- **Molecular weight, pI**
- **Comment, literature citations**
- **Species**

HumanCyc -- HumanCyc.org

- **Derived from Ensembl and LocusLink**
- **Tier 2 PGDB**
- **Curation has just resumed**

- **235 metabolic pathways**
- **1,523 small-molecule reactions**
- **1,188 substrates**

- ***Genome Biology* 6:1-17 2004.**

EcoCyc Project – EcoCyc.org

- ***E. coli* Encyclopedia**

- Review-level Model-Organism Database for *E. coli*
- Tracks evolving annotation of the *E. coli* genome and cellular networks
- The two paradigms of EcoCyc

- **Collaborative development via Internet**

- Paulsen (TIGR) – Transport, flagella, DNA repair
- Collado (UNAM) -- Regulation of gene expression
- Keseler, Shearer (SRI) -- Metabolic pathways, cell division, proteases
- Karp (SRI) -- Bioinformatics

Paradigm 1:

EcoCyc as Textual Review Article

- **All gene products for which experimental literature exists are curated with a minireview summary**
 - Found on protein and RNA pages, not gene pages!
 - 3257 gene products contain summaries
- **Summaries cover function, interactions, mutant phenotypes, crystal structures, regulation, and more**
- **Additional summaries found in pages for operons, pathways**
- **EcoCyc cites 14,269 publications**

Paradigm 2:

EcoCyc as Computational Symbolic Theory

- **Highly structured, high-fidelity knowledge representation provides computable information**
- **Each molecular species defined as a DB object**
 - Genes, proteins, small molecules
- **Each molecular interaction defined as a DB object**
 - Metabolic reactions
 - Transport reactions
 - Transcriptional regulation of gene expression
- **220 database fields capture extensive properties and relationships**

Demonstration

Pathway Tools Schema and Semantic Inference Layer

Guiding Principles for the Pathway Tools Ontology of Biological Function

- **Encode distinct molecular species as separate objects**
- **Describe all molecular interactions as reactions**
- **Layered approach:**
 - Molecular species form the base
 - Reactions built from molecular species
 - Pathways built from reactions
- **Link catalyst to reaction via Enzymatic-Reaction**



Pathway Tools Ontology / Schema

- **Ontology classes: 1621**
 - Datatype classes: Define objects from genomes to pathways
 - Classification systems / controlled vocabularies
 - ◆ Pathways, chemical compounds, enzymatic reactions (EC system)
 - ◆ Protein Feature ontology
 - ◆ Cell Component Ontology
 - ◆ Evidence Ontology
- **Comprehensive set of 279 attributes and relationships**

Overview of Schema Presentation

- **Survey of important classes**
- **What slots are present within these classes**
- **How objects are linked together to form a network**

Use GKB Editor to Inspect the Pathway Tools Ontology

- **GKB Editor = Generic Knowledge Base Editor**
- **Type in Navigator window: (GKB) or**
- **[Right-Click] Edit->Ontology Editor**

- **View->Browse Class Hierarchy**
- **[Middle-Click] to expand hierarchy**
- **To view classes or instances, select them and:**
 - **Frame -> List Frame Contents**
 - **Frame -> Edit Frame**

Root Classes in the Pathway Tools Ontology

- **Chemicals** -- All molecules
- **Polymer-Segments** -- Regions of polymers
- **Protein-Features** -- Features on proteins
- **Paralogous-Gene-Groups**

- **Organisms**

- **Generalized-Reactions** -- Reactions and pathways
- **Enzymatic-Reactions** -- Link enzymes to reactions they catalyze
- **Regulation** -- Regulatory interactions

- **CCO** -- Cell Component Ontology
- **Evidence** -- Evidence ontology

- **Notes** -- Timestamped, person-stamped notes
- **Organizations**
- **People**
- **Publications**

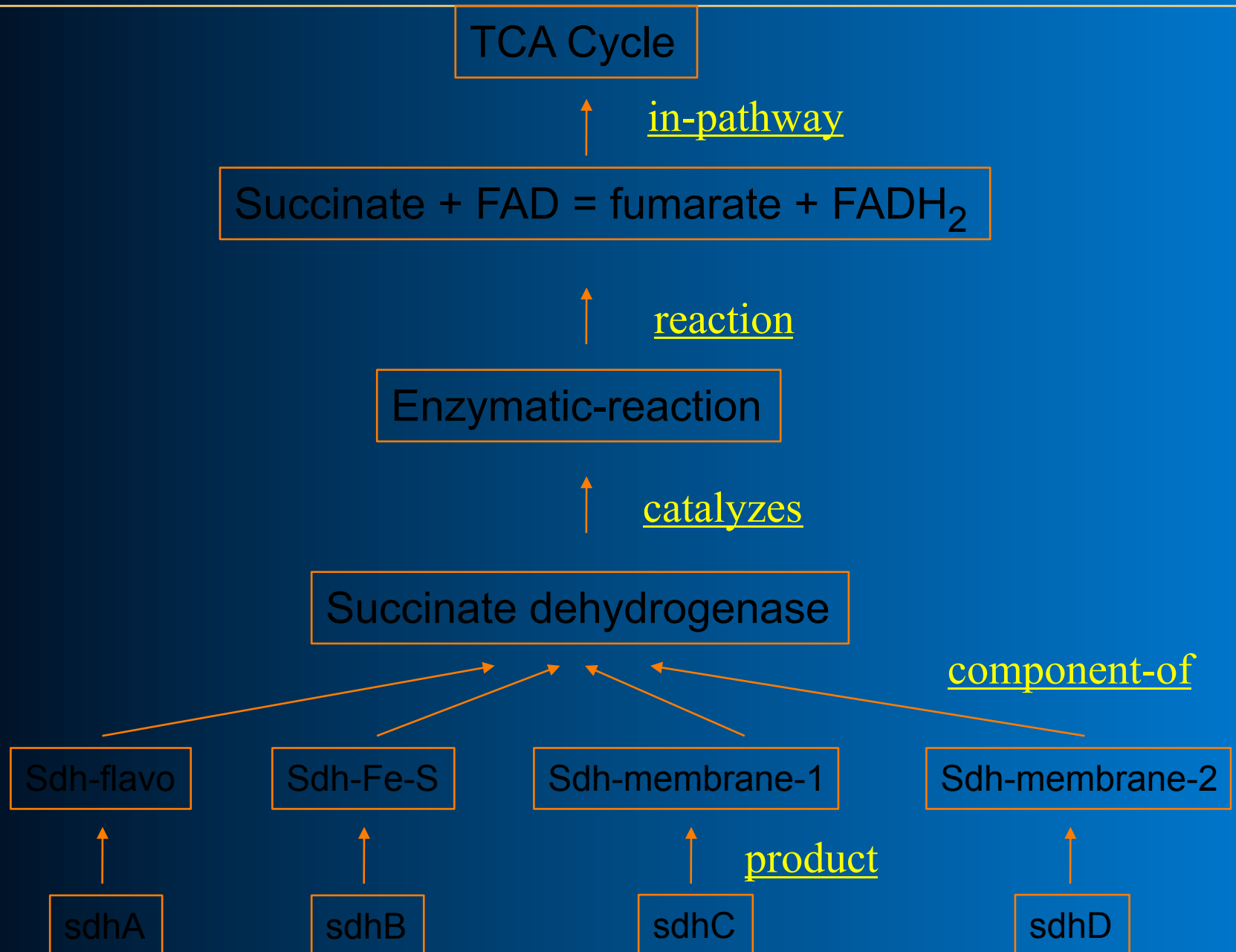
Principal Classes

- **Class names are usually capitalized, plural, separated by dashes**
- **Genetic-Elements, with subclasses:**
 - Chromosomes
 - Plasmids
- **Genes**
- **Transcription-Units**
- **RNAs**
 - rRNAs, snRNAs, tRNAs, Charged-tRNAs
- **Proteins, with subclasses:**
 - Polypeptides
 - Protein-Complexes

Principal Classes

- **Reactions**
- **Enzymatic-Reactions**
- **Pathways**
- **Compounds-And-Elements**
- **Regulation**

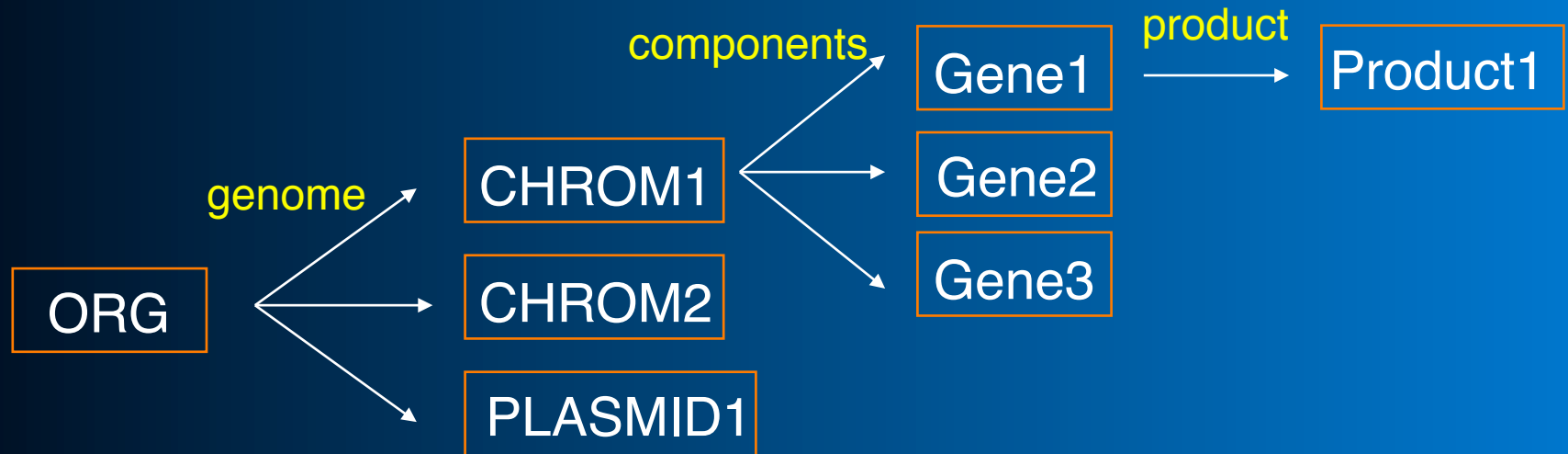
Semantic Network Diagrams



Pathway Tools Schema and Semantic Inference Layer

*Genes, Operons, and
Replicons*

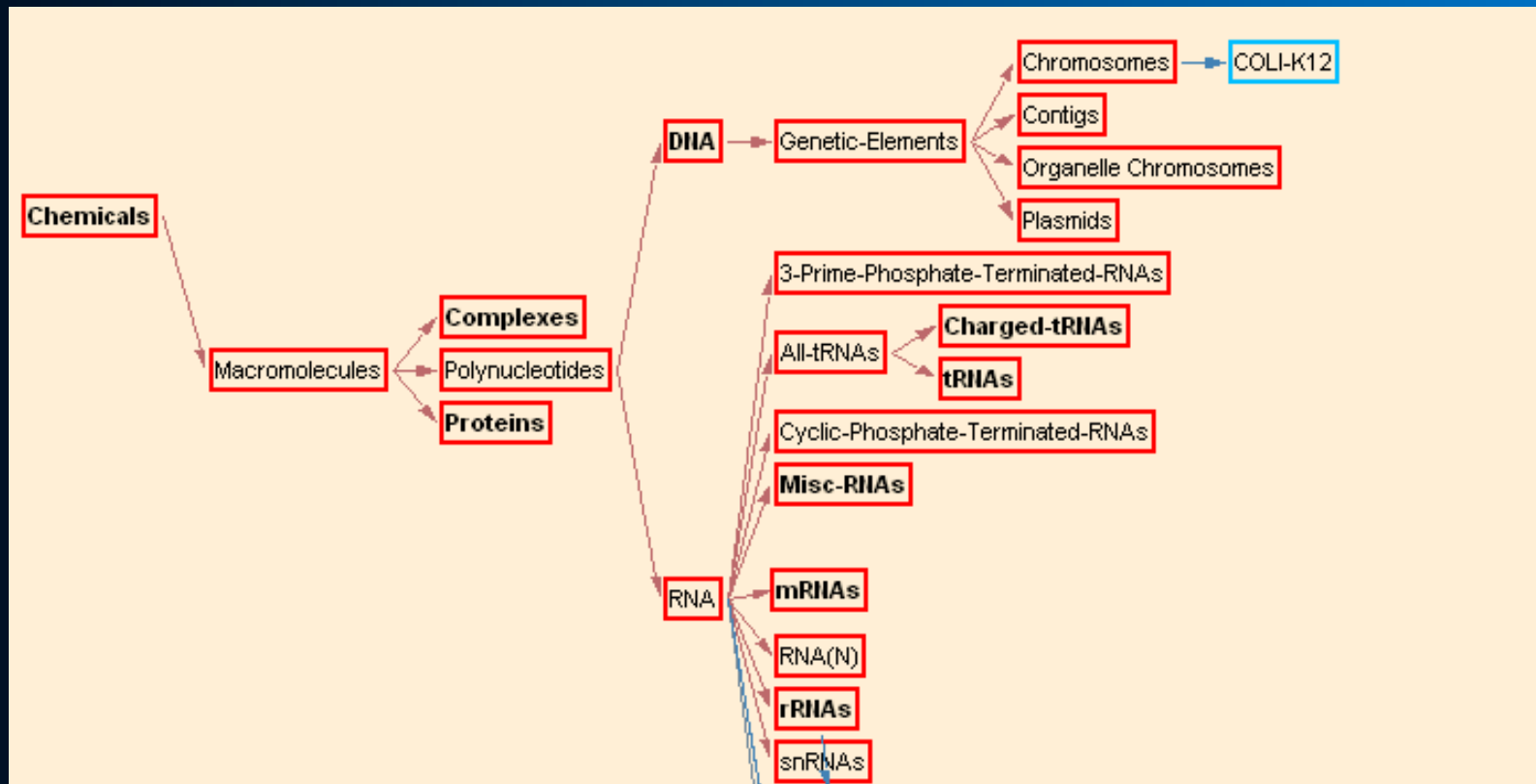
Representing a Genome



- **Classes:**

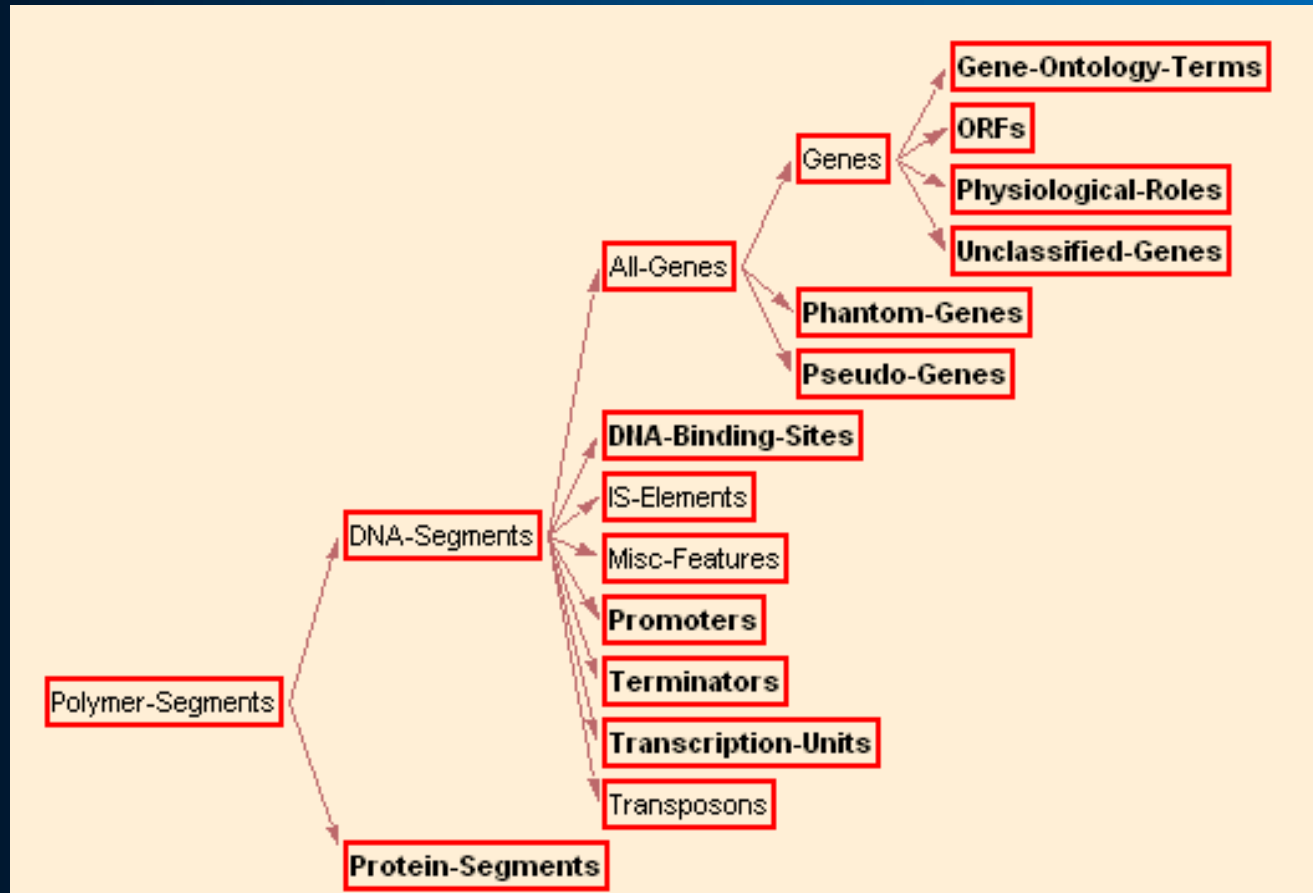
- ORG is of class Organisms
- CHROM1 is of class Chromosomes
- PLASMID1 is of class Plasmids
- Gene1 is of class Genes
- Product1 is of class Polypeptides or RNA

Polynucleotides



Review slots of COLI and of COLI-K12

Polymer-Segments



Review slots of Genes

Proteins

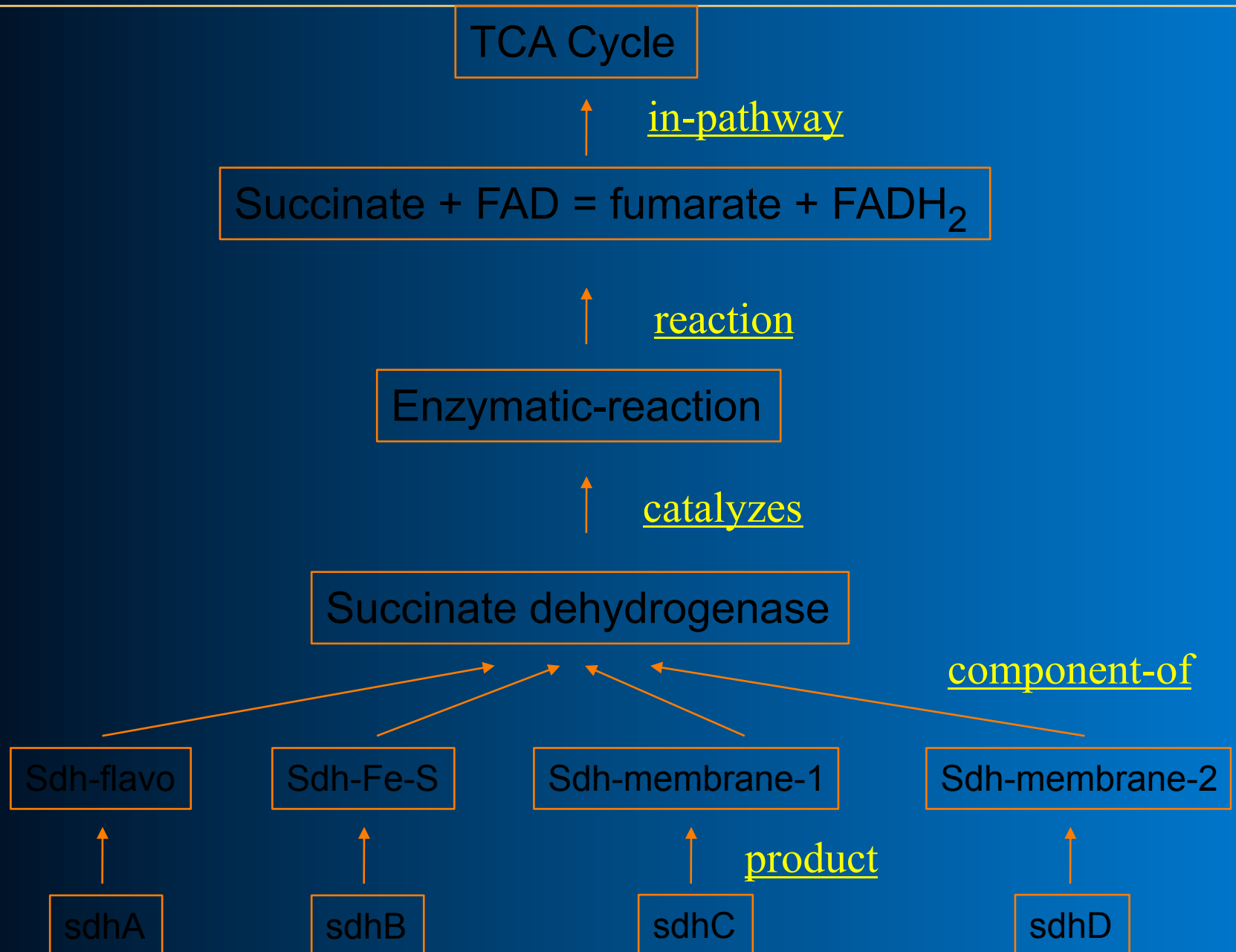
Proteins and Protein Complexes

- **Polypeptide: the monomer protein product of a gene (may have multiple isoforms, as indicated at gene level)**
- **Protein complex: proteins consisting of multiple polypeptides or protein complexes**
- **Example: DNA pol III**
 - DnaE is a polypeptide
 - pol III core enzyme contains DnaE, DnaQ, HolE
 - pol III holoenzyme contains pol III core enzyme plus three other complexes

Slots of Proteins (DnaE)

- **comments, citations**
- **pl, molecular-weight**
- **features**
- **component-of**
- **gene**
- **catalyzes** [link to Enzymatic-Reaction]
- **dblinks**

Semantic Network Diagrams



Semantic Inference Layer

- **Reactions-of-protein (prot)**
 - Returns a list of rxns this protein catalyzes
- **Transcription-units-of-proteins(prot)**
 - Returns a list of TU's activated/inhibited by the given protein
- **Transporter? (prot)**
 - Is this protein a transporter?
- **Polypeptide-or-homomultimer?(prot)**
- **Transcription-factor? (prot)**
- **Obtain-protein-stats**
 - Returns 5 values
 - ◆ Length of : all-polypeptides, complexes, transporters, enzymes, etc...

***Compounds / Reactions /
Pathways***

Compounds / Reactions / Pathways

- **Think of a three tiered structure:**
 - Compounds at the bottom
 - Reactions built on top of compounds
 - Pathways built on top of reactions
- **Metabolic network can be defined by reactions alone**
- **Pathways are an additional “optional” structure**
- **Some reactions not part of a pathway**
- **Some reactions have no attached enzyme**
- **Some enzymes have no attached gene**

Compounds

Slots of Compounds

- **common-name, abbrev-name, synonyms**
- **comment, citations**
- **charge, gibbs-0, molecular-weight**
- **empirical-formula**
- **structure-atoms, structure-bonds**
- **appears-in-left-side-of, appears-in-right-side-of**

Semantic Inference Layer

- **Reactions-of-compound (cpd)**
- **Pathways-of-compound (cpd)**
- **Activated/inhibited-by? (cpds slots)**
 - Returns a list of enrxtns for which a cpd in cpds is a modulator (example slots: activators-all, activators-allosteric)
- **All-substrates (rxns)**
 - All unique substrates specified in the given rxns
- **Has-structure-p (cpd)**

Reactions

Reactions

- **Represent information about a reaction that is independent of enzymes that catalyze the reaction**
- **Connected to enzyme(s) via enzymatic reaction frames**
- **Classified with EC system when possible**
- **Example: 2.7.7.7 – DNA-directed DNA polymerization**
 - Carried out by five enzymes in *E. coli*

Slots of Reaction Frames

- **Keq**
- **Left and Right (reactants / products)**
 - Can include modified forms of proteins, RNAs, etc here
- **Enzymatic-reaction**
- **In-pathway**

Semantic Inference Layer

- **Genes-of-reaction (rxn)**
- **Substrates-of-reaction (rxn)**
- **Enzymes-of-reaction (rxn)**
- **Lacking-ec-number (organism)**
 - Returns list of rxns with no ec numbers in that database
- **Get-reaction-direction-in-pathway (pwy rxn)**
- **Reaction-type(rxn)**
 - ◆ Indicates types of Rxn as: Small molecule rxn, transport rxn, protein-small-molecule rxn (one substrate is protein and one is a small molecule), protein rxn (all substrates are proteins), etc.
- **All-rxns(type)**
 - Specify the type of reaction (see above for type)
- **Obtain-rxn-stats**
 - Returns six values
 - ◆ Length of : all-rxns, transport, non-transport, etc...

Enzymatic Reactions (DnaE and 2.7.7.7)

- **A necessary bridge between enzymes and “generic” versions of reactions**
- **Carry information specific to an enzyme/reaction combination:**
 - Cofactors and prosthetic groups
 - Alternative substrates
 - Links to regulatory interactions
- **Frame is generated when protein is associated with reaction (via protein or reaction editor)**

Regulation of Enzyme Activity

Frame Editor

Application Knowledge Base Frame Value Slot Preferences

Editing Instance REG0-6616, instance of Regulation-of-Enzyme-Activity

CITATIONS — "11515538"

COMMENT

COMMENT-INTERNAL

COMMON-NAME

CREATION-DATE — 16-Jun-2007 01:08:08

CREATOR — paley

CREDITS

DATA-SOURCE

HISTORY

MECHANISM — NONCOMPETITIVE

MODE — "."

PHYSIOLOGICALLY-RELEVANT?

REGULATED-BY

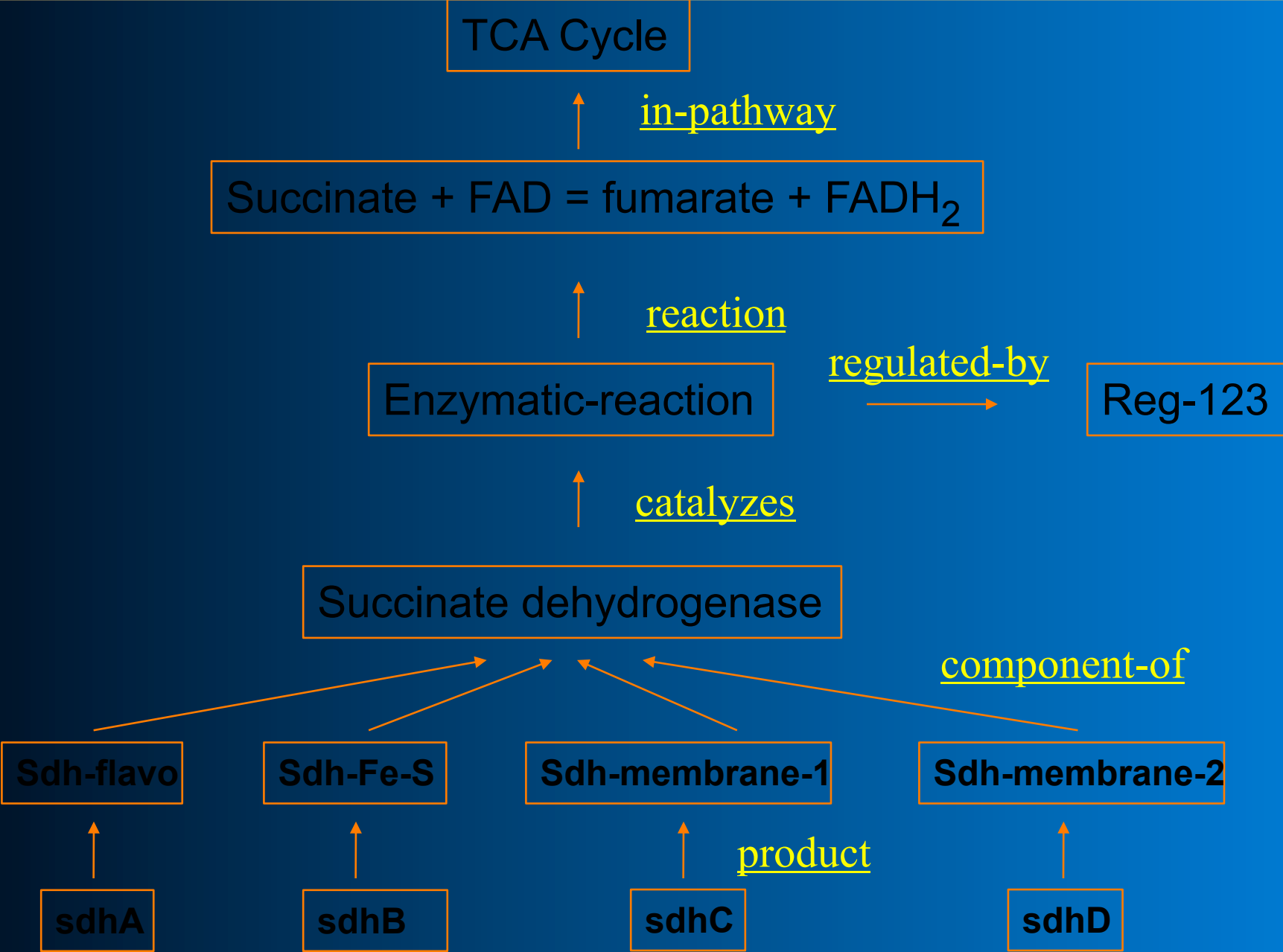
REGULATED-ENTITY — aminopeptidase

REGULATOR — EDTA

SCHEMA? — T
inherited from Regulation

SYNONYMS

Semantic Network Diagrams



***Pathway Tools Schema and Semantic
Inference Layer: Pathways***

Pathway Ontology

- **Slots in pathway:**

- Reaction-List, Predecessor-List



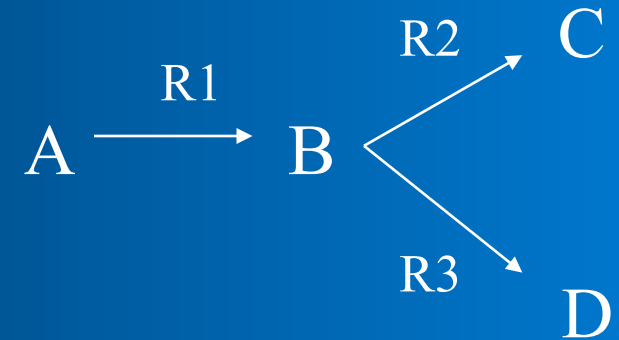
R1: Left = A, Right = B

R2: Left = B, Right = C

R3: Left = C, Right = D

Predecessor list:

(R1 R2) (R2 R3)



R1: Left = A, Right = B

R2: Left = B, Right = C

R3: Left = B, Right = D

Predecessor list:

(R1 R2) (R1 R3)

Super-Pathways

- **Collection of pathways that connect to each other via common substrates or reactions, or as part of some larger logical unit**
- **Can contain both sub-pathways and additional connecting reactions**
- **Can be nested arbitrarily**
- **REACTION-LIST: a pathway ID instead of a reaction ID in this slot means include all reactions from the specified pathway**
- **PREDECESSORS: a pathway ID instead of a tuple in this slot means include all predecessor tuples from the specified pathway**

Querying Pathways Programmatically

- See <http://bioinformatics.ai.sri.com/ptools/ptools-resources.html>
- **(all-pathways)**
- **(base-pathways)**
 - Returns list of all pathways that are not super-pathways
- **(genes-of-pathway pwy)**
- **(unique-genes-of-pathway pwy)**
 - Returns list of all genes of a pathway that are not also part of other pathways
- **(enzymes-of-pathway pwy)**
- **(substrates-of-pathway pwy)**
- **(variants-of-pathway pwy)**
 - Returns all pathways in the same variant class as a pathway
- **(get-predecessors rxn pwy), (get-successors rxn pwy)**
- **(get-rxn-direction-in-pathway pwy rxn)**
- **(pathway-inputs pwy), (pathway-outputs pwy)**
 - Returns all compounds consumed (produced) but not produced (consumed) by pathway (ignores stoichiometry)

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**

Regulation in 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**

Regulation

- **Class Regulation with subclasses that describe different biochemical mechanisms of regulation**
- **Slots:**
 - Regulator
 - Regulated-Entity
 - Mode
 - Mechanism

Regulation of Enzyme Activity

- **Class Regulation-of-Enzyme-Activity**
- **Each instance of the class describes one regulatory interaction**
- **Slots:**
 - Regulator -- usually a small molecule
 - Regulated-Entity -- an Enzymatic-Reaction
 - Mechanism -- One of:
 - ◆ Competitive, Uncompetitive, Noncompetitive, Irreversible, Allosteric, Unkmech, Other
 - Mode -- One of: + , -

Transcription Initiation

- **Class Regulation-of-Transcription-Initiation**

- **Slots:**

- Regulator -- instance of Proteins or Complexes (a transcription-factor)
- Regulated-Entity -- instance of Promoters or Transcription-Units or Genes
- Mode -- One of: + , -

Other Features of Ontology

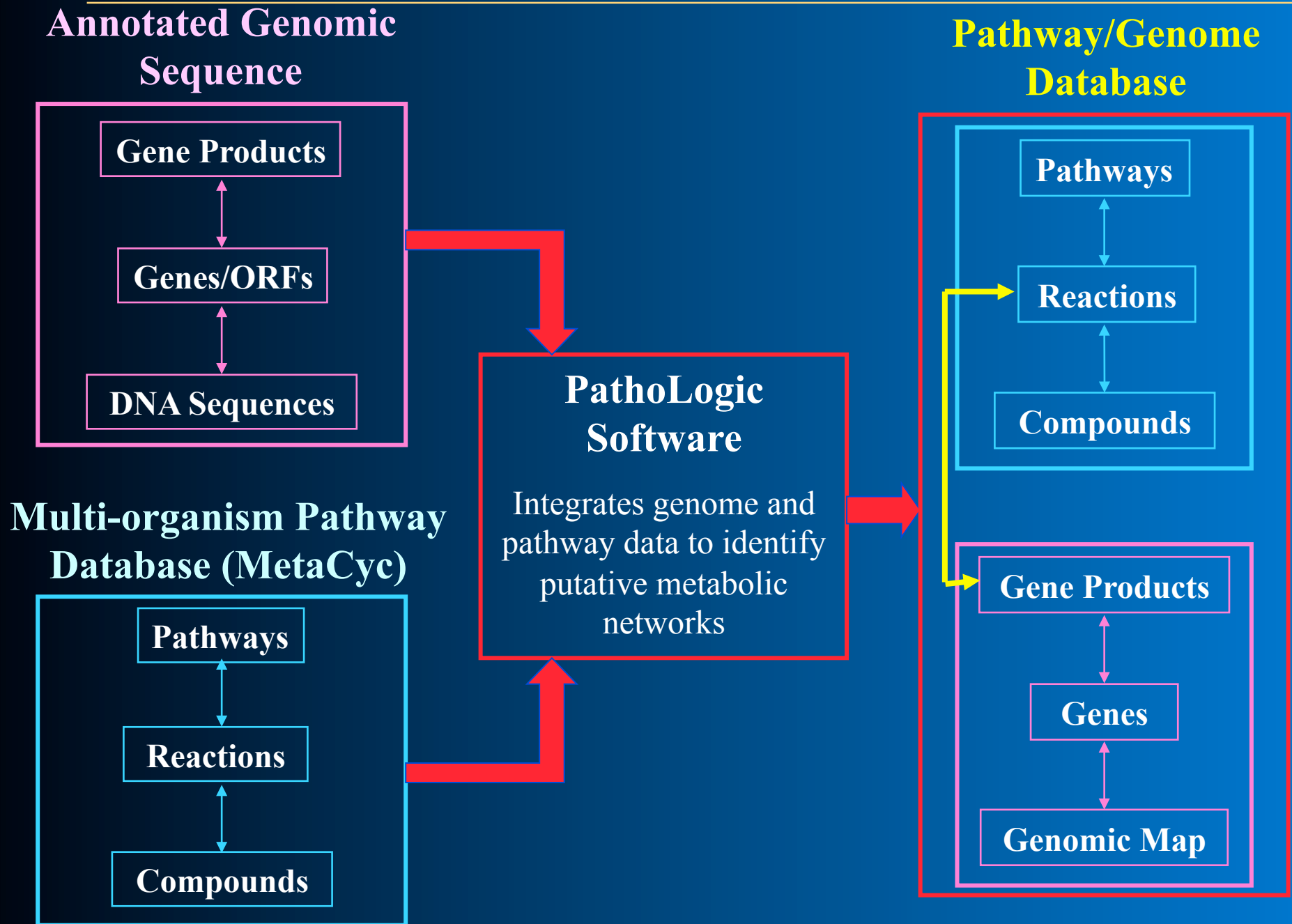
- Evidence codes
- Curator crediting system

Inference Algorithms

PathoLogic: Inference of Pathway Complement

- **An additional level of inference after genome annotation**
- **Place predicted genes in their biochemical context**
- **Information reduction device**
- **Assess coherence of the set of genes in a genome**
- **Identify pathway holes and singleton enzymes**
- **Provides a framework for analysis of functional-genomics data**

Inference of Metabolic Pathways



Genbank Format :

```
gene          422054..423490
              /gene="aroE"
CDS           422054..423490
              /gene="aroE"
              /label="CT370"
              /product="Shikimate 5-Dehydrogenase"
              /db_xref="PID:g3328794"
```

PathoLogic Format :

```
ID           CT370
NAME         aroE
STARTBASE   422054
ENDBASE     423490
PRODUCT     Shikimate 5-Dehydrogenase
DBLINK      PID:g3328794
PRODUCT-TYPE P
EC          1.1.1.25
```

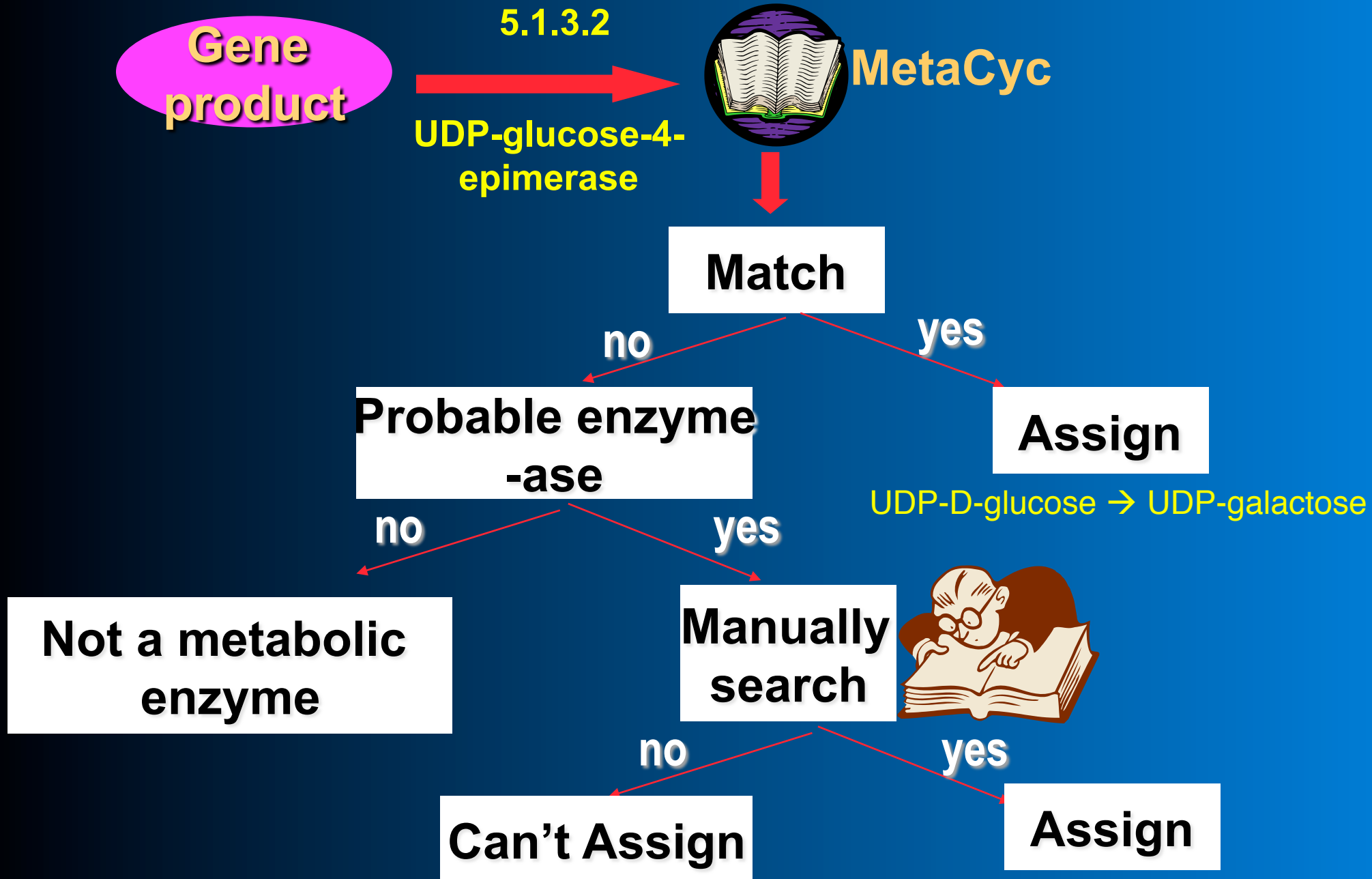
Pathway Prediction

- **Step 1: Infer reactome**
- **Step 2: Infer metabolic pathways from reactome**

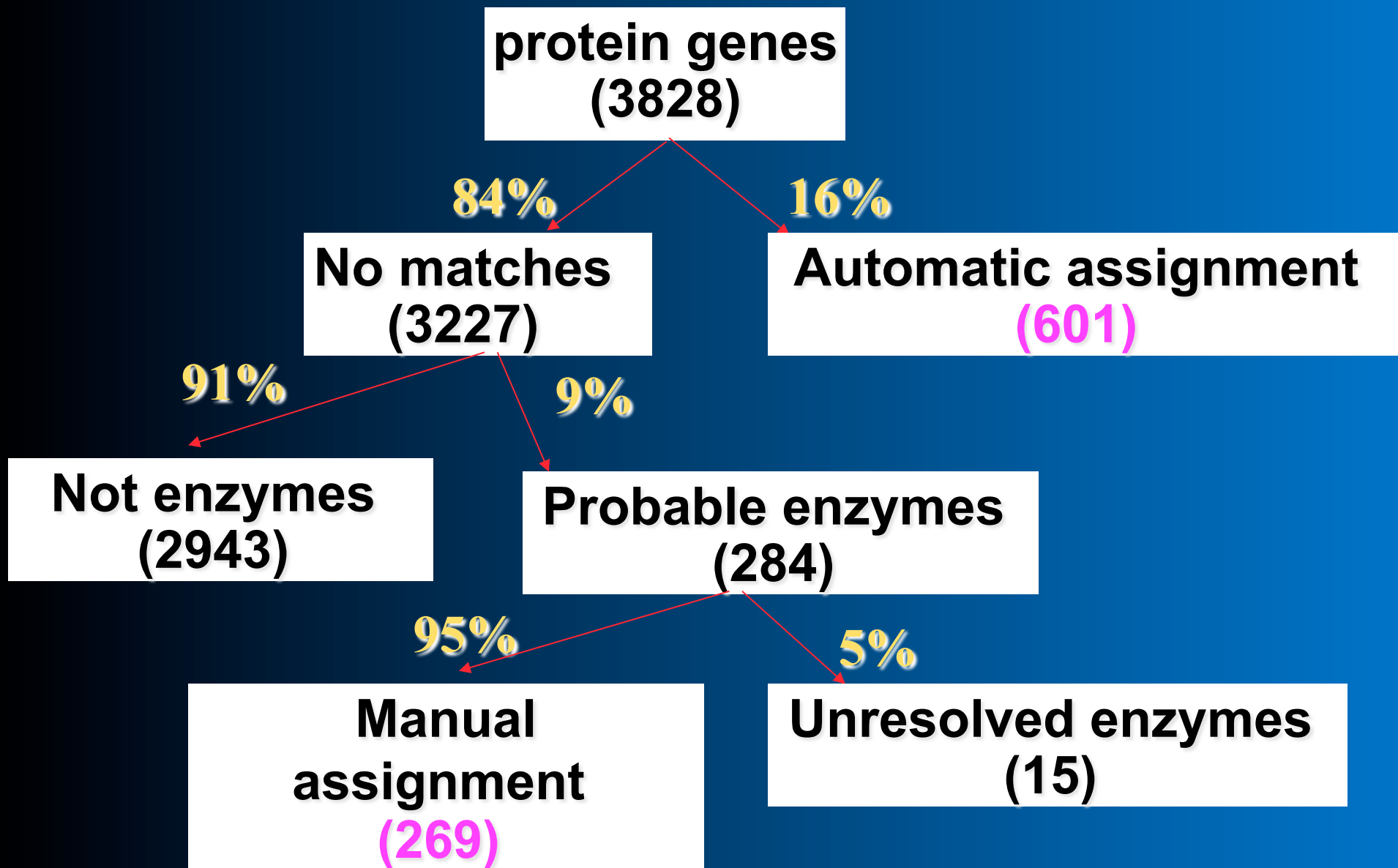
Inference of Reactome

- **Given genome annotation, infer metabolic reactions that can be catalyzed by the genome**
 - EC numbers
 - Enzyme names
 - Gene Ontology annotations
- **Complications:**
 - Most genomes contain a subset of above annotations
 - Enzyme names sometimes ambiguous
 - Some reactions occur in multiple pathways
 - ◆ 99 of 744 reactions in *E. coli*
 - Pathway variants

Match Enzymes to Reactions



Vibrio cholerae Enzyme Matching Results



Pathway Prediction Algorithm

- **Two pathway lists:**
 - U: Undecided status
 - K: Keep
- **Initialize U to contain all MetaCyc pathways for which at least one reaction has an enzyme**

Pathway Prediction Algorithm





- **For each P in U:**
 - If current organism is outside taxonomic range of P AND at least one reaction in P lacks an enzyme, delete P from U
 - If all reactions of P designated as key reactions have no enzyme, delete P from U

Pathway Prediction Algorithm

- **Iterate through P in U until U is unchanged:**
 - If P should be kept, move P to K
 - ◆ A reaction in P is unique to P and has an enzyme
 - ◆ At most one reaction in P has no enzyme
 - ◆ The enzymes present for P are not a subset of the enzymes present for a variant pathway of P
 - If P should be deleted, delete P from U
 - ◆ At most one reaction R in P has an enzyme, and R is not unique to P
 - ◆ The pathway is a biosynthetic pathway missing its final steps
 - ◆ The pathway is a catabolic pathway missing its initial steps
- **Accuracy: 91%**

Pathway Evidence Report

Biosynthesis: Cofactors, Prosthetic Groups, Electron Carriers

Pathway	Pathway Glyph	Total Rxns	Rxns Present in <i>V. cholerae</i>	Rxns Present in Other Pwys	Other Pwys
biosynthesis of proto- and siroheme		15	12	2	tRNA charging pathway cobalamin biosynthesis II (aerobic pathway)
biotin biosynthesis I		4	4	0	(none)
cobalamin biosynthesis I		7	6	2	cobalamin biosynthesis II (aerobic pathway)
cobalamin biosynthesis II (aerobic pathway)		18	5	3	cobalamin biosynthesis I biosynthesis of proto- and siroheme

Limitations of Pathway Inference

- **Can be misled by missing or incorrect functional assignments**
- **No sequences known for many enzymes**
- **Uncertainty for short pathways**

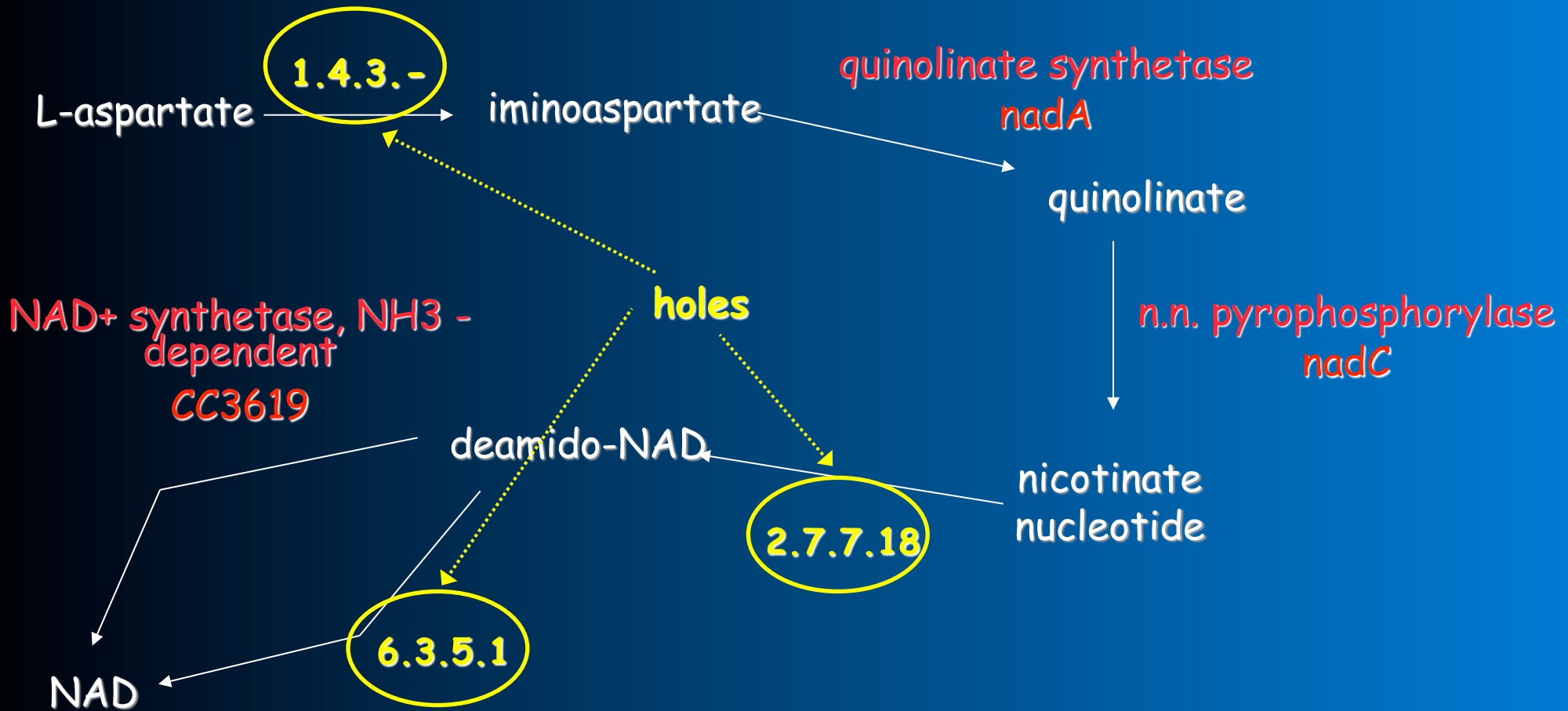
-
- **In 90 minutes, I got to here**
 - **Included a 10-15 min demo**
 - **3/10/2010 Brutlag class lecture**

-
- **Hole filler 10**
 - **Forward prop 10**
 - **Comp analysis 10**
 - **Choke points 5**

- **Groups**
- **Overviews**
- **Omics viewers**









Pathway Hole Filling

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



**Step 1: collect query
isozymes of function A
based on EC#**

**Step 2: BLAST
against target
genome**

- organism 1 enzyme A* 
- organism 2 enzyme A* 
- organism 3 enzyme A* 
- organism 4 enzyme A* 
- organism 5 enzyme A* 
- organism 6 enzyme A* 
- organism 7 enzyme A* 
- organism 8 enzyme A* 

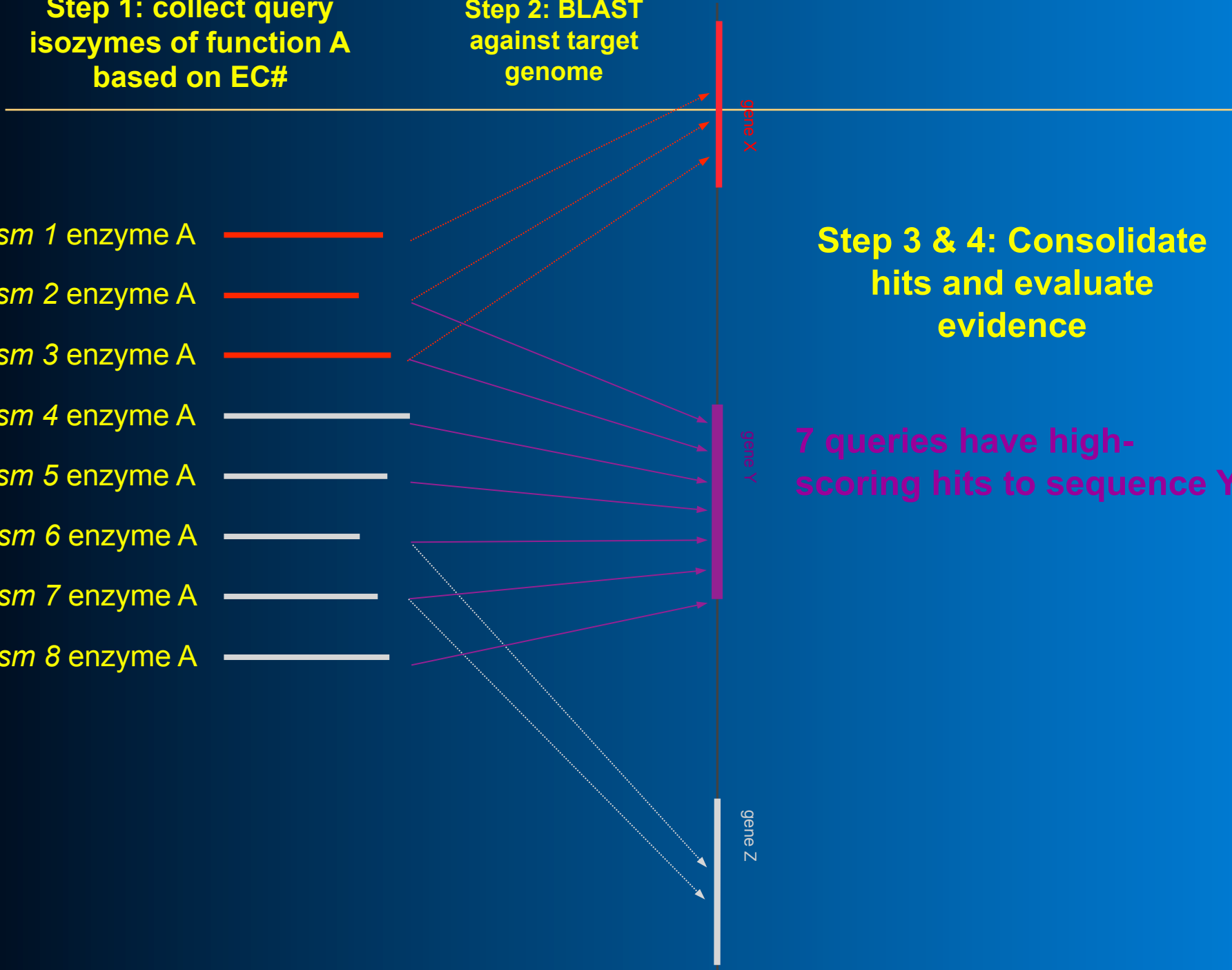
gene X

gene Y

gene Z

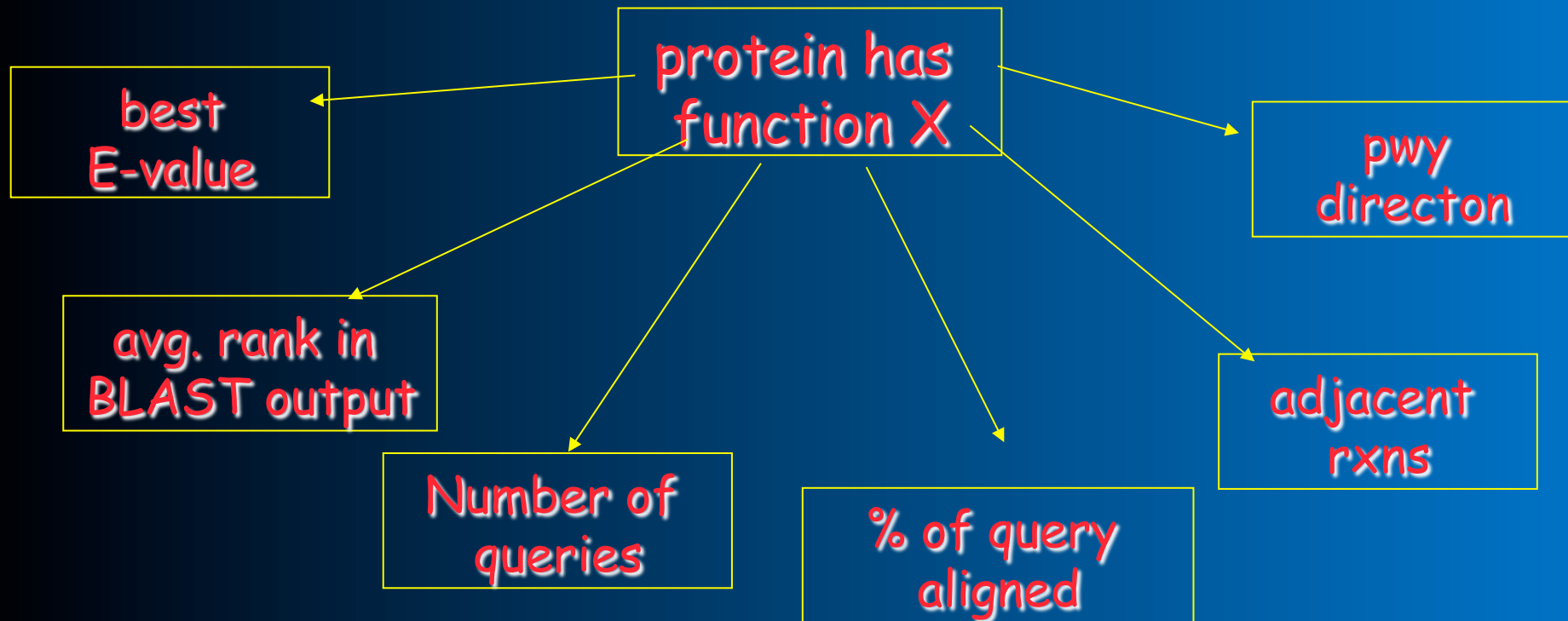
**Step 3 & 4: Consolidate
hits and evaluate
evidence**

**7 queries have high-
scoring hits to sequence Y**



Bayes Classifier

$P(\text{protein has function } X |$
E-value, avg. rank, aln. length, etc.)



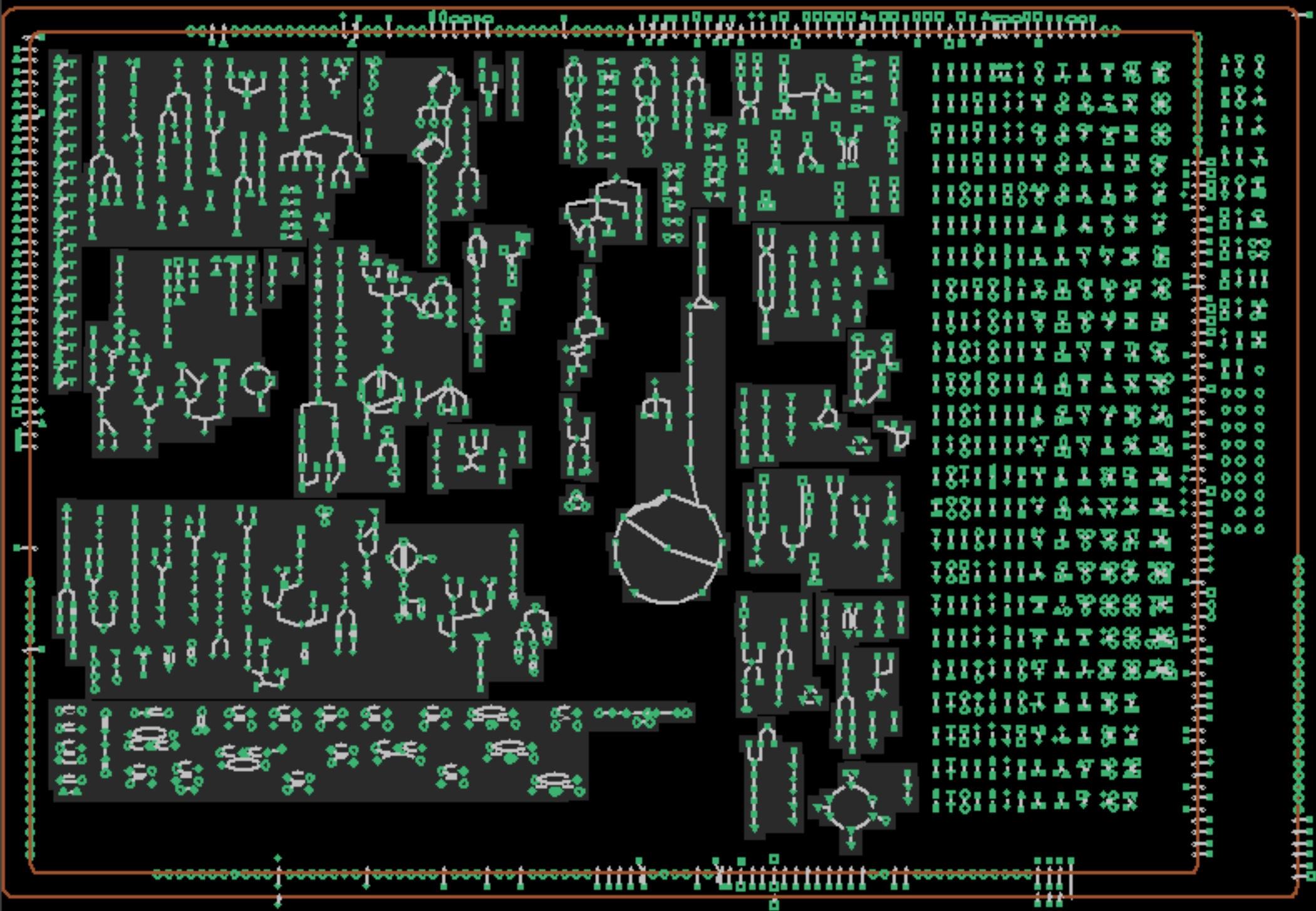
Pathway Hole Filler

- **Why should hole filler find things beyond the original genome annotation?**
- **Reverse BLAST searches more sensitive**
- **Reverse BLAST searches find second domains**
- **Integration of multiple evidence types**

PathoLogic Step 6: Build Cellular Overview Diagram

- **Diagram encompassing metabolic, transport, and other cellular networks**
- **Automatically generated for every BioCyc DB using advanced graph layout algorithm**
- **Harness the power of the human visual system to interpret patterns in a mechanistic context**
- **Can be zoomed, interrogated, and painted with experimental or comparative data**

E. coli K-12 Cellular Overview



Escherichia coli

Home

Back

Forward

History

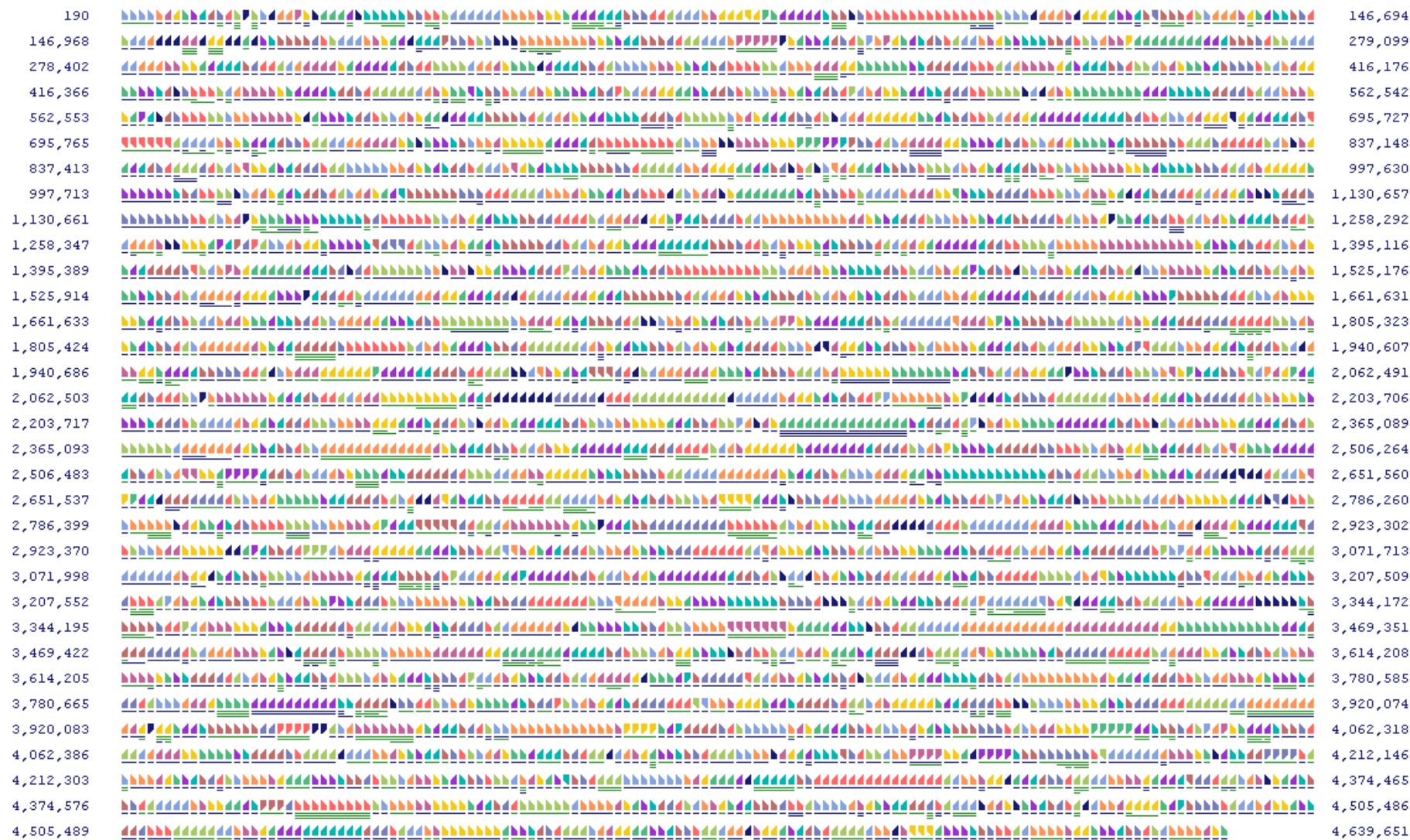
Next Answer

Clone

Save DB

E. coli Chromosome Global Overview

Escherichia coli K-12 Chromosome:

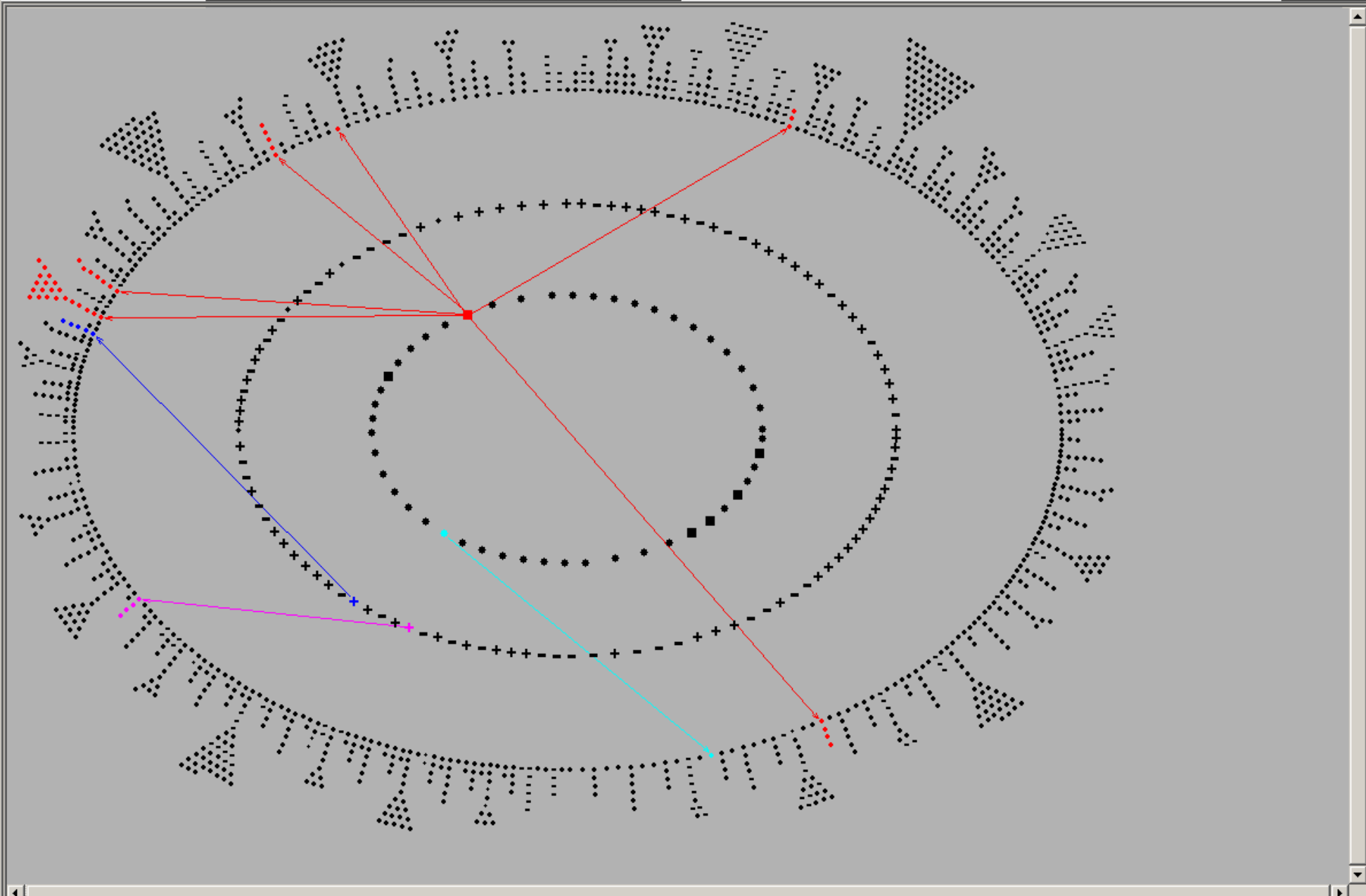


Command: :Cmd Menu Node Network "chbR"

Command: :Cmd Menu Node Network "prpR"

Command: Genbro Show Chromosome Global View

Command: []



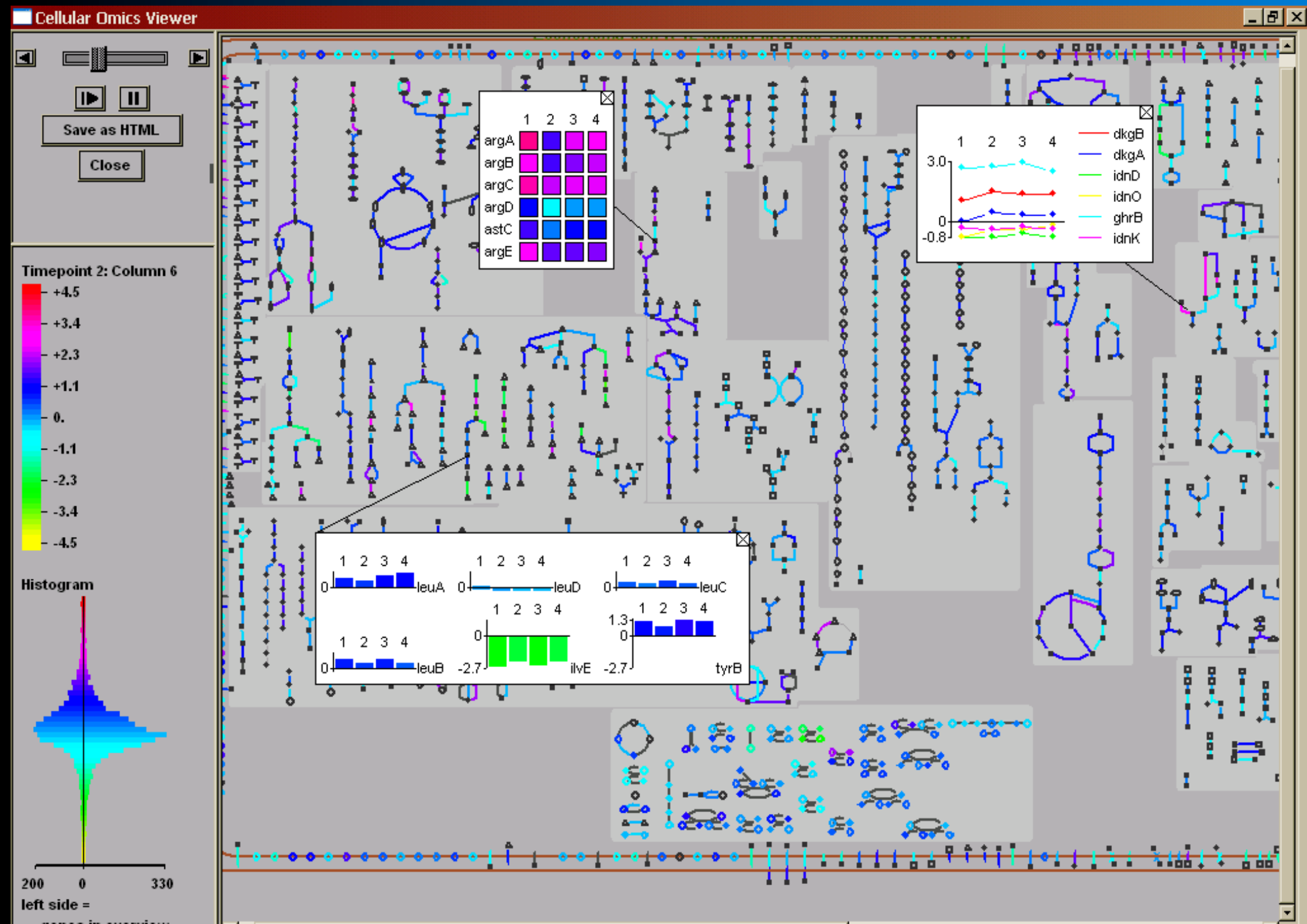
Command: :Cmd Menu Node Network "uhpA"

Command: :Cmd Menu Node Network "chbR"

Command: :Cmd Menu Node Network "prpR"

Command: []

Omics Data Graphing

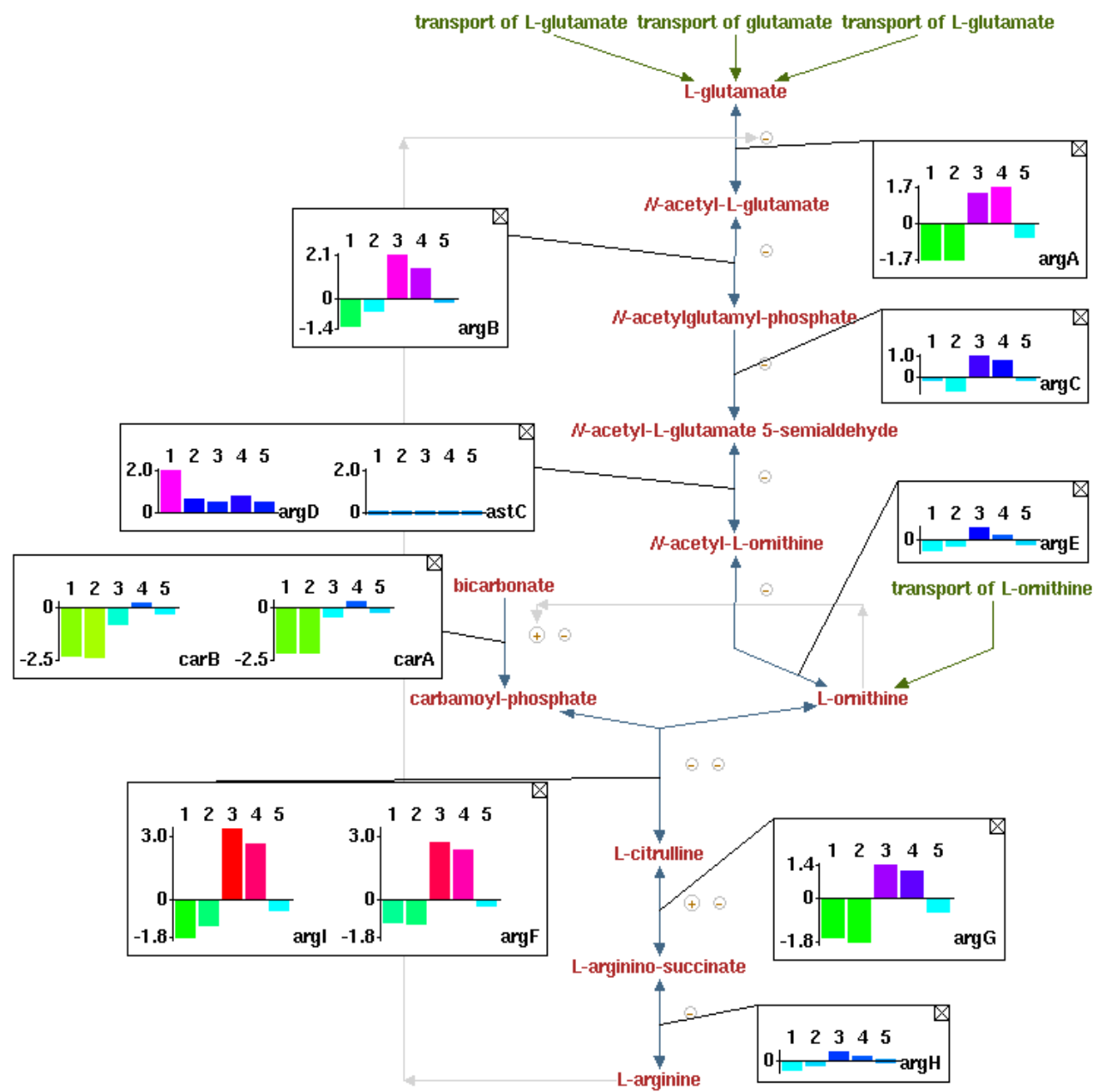


Escherichia coli K-12 substr. MG1655

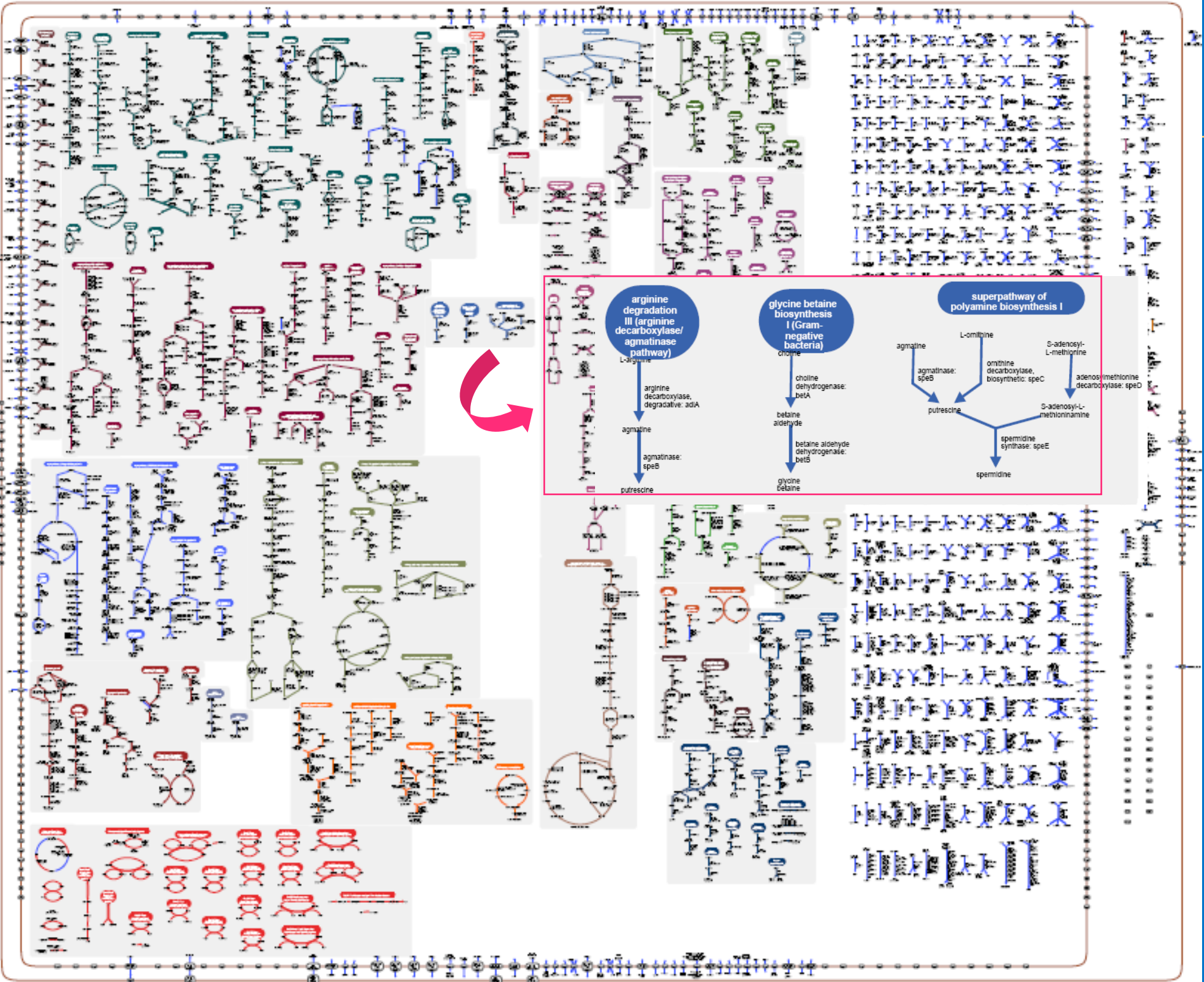
Home Back Forward History Next Answer Clone Save DB

Escherichia coli K-12 substr. MG1655 Pathway: arginine biosynthesis I

More Detail Less Detail Species Comparison



EcoCyc: Escherichia coli K-12 Cellular Overview



Genome Poster



Symbolic Systems Biology

Definition:

Global analyses of biological systems using symbolic computing

Symbolic Systems Biology

- **“Symbolic computing is concerned with the representation and manipulation of information in symbolic form. It is often contrasted with numeric representation.” -- R. Cameron**
- **Examples of symbolic computation:**
 - Symbolic algebra programs, e.g., Mathematica, Graphing Calculator
 - Compilers and interpreters for programming languages
 - Database query languages
 - Text analysis programs, e.g., Google
 - String matching for DNA and protein sequences
 - Artificial Intelligence methods, e.g., expert systems, symbolic logic, machine learning, natural language understanding

Symbolic Systems Biology

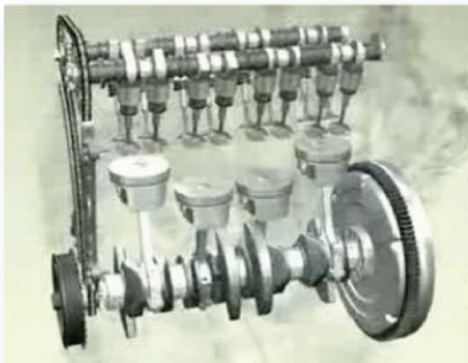
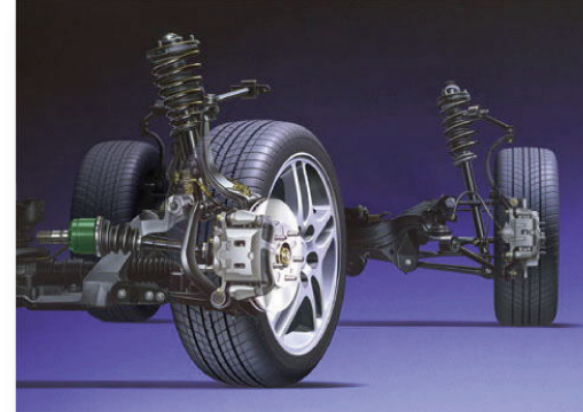
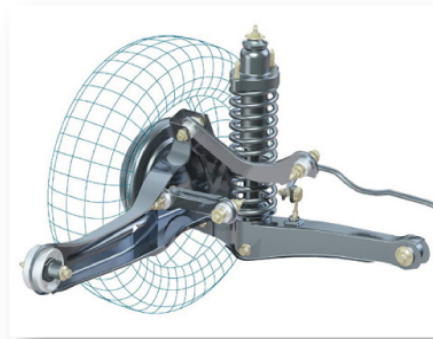
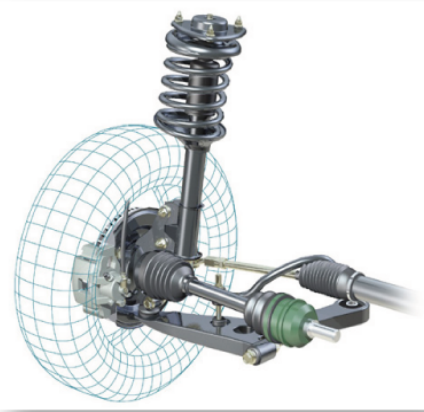
- **Concerned with different questions than quantitative systems biology**
- **Symbolic analyses can in many cases produce answers when quantitative approaches fail because of lack of parameters or intractable mathematics**
- **Symbolic computation is intimately dependent on the use of structured ontologies**

Symbolic Computation on PGDBs: Complex Queries

- **Show metabolic enzymes regulated by a specified transcription factor**
- **For transcription factor F:**
- **Find all promoters F regulates**
 - Find all genes in the operons controlled by those promoters
 - ◆ Find their protein products
 - Find the reactions they catalyze
 - » Highlight them in the diagram

Critiquing the Parts List

Slide thanks to Hirotada Mori
(minus the banana!)



Transport Inference Parser

- **Problem: Compare the transportable substrates of an organism with the metabolic reactions of the organism**
- **Sub-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:**

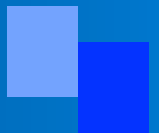


Dead End Metabolites

- **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

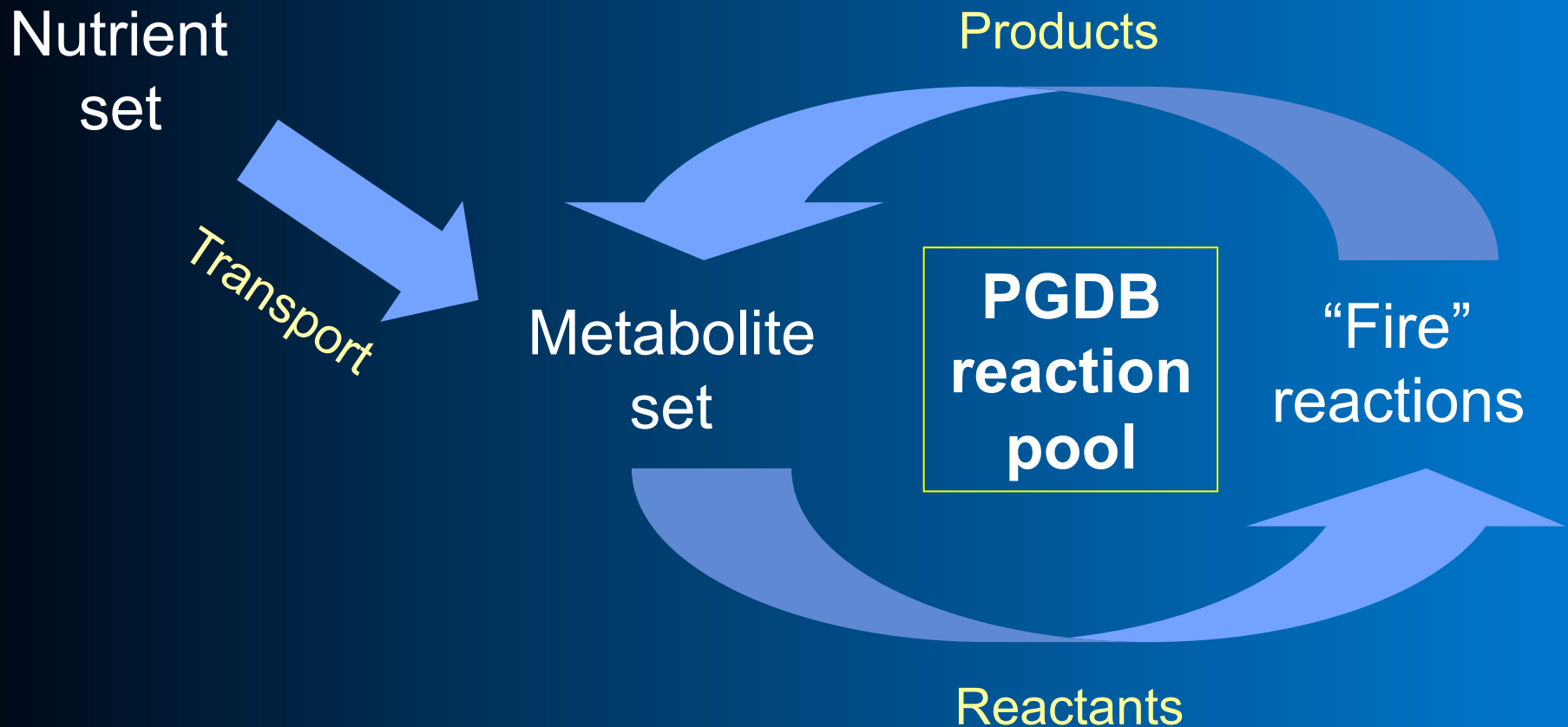
Reachability Analysis of Metabolic Networks

- **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



Algorithm: Forward Propagation Through Production System

- Each reaction becomes a production rule
- Each metabolite in nutrient set becomes an axiom



Starting Nutrients: A, B, C, E, F



Produced Compounds: W, Y, Z

Starting Nutrients: A, B, C, E, F



- Need to supply some starting G
- But: G is regenerated; cells will likely contain some small amount of G

Initial Metabolite Nutrient Set

(Total: 21 compounds)

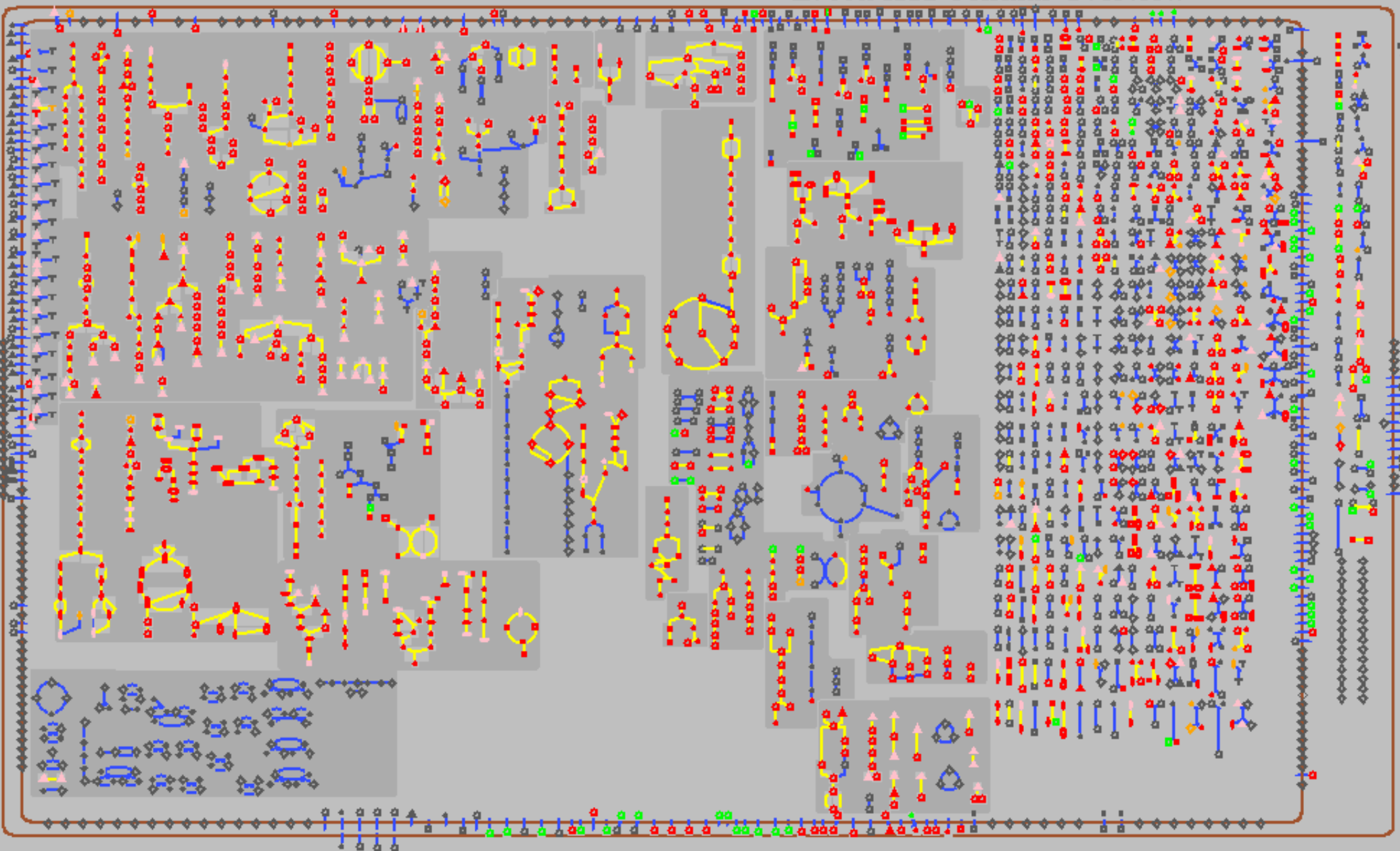
<p>Nutrients (8) (M61 Minimal growth medium)</p>	<p>H^+, Fe^{2+}, Mg^{2+}, K^+, NH_3, SO_4^{2-}, PO_4^{2-}, Glucose</p>
<p>Nutrients (10) (Environment)</p>	<p>Water, Oxygen, Trace elements (Mn^{2+}, Co^{2+}, Mo^{2+}, Ca^{2+}, Zn^{2+}, Cd^{2+}, Ni^{2+}, Cu^{2+})</p>
<p>Bootstrap Compounds (3)</p>	<p>ATP,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

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

- **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)
 - ◆ $A \rightarrow B$ $B' \rightarrow C$
- Incomplete knowledge of *E. coli* metabolic network
 - ◆ $A + B \rightarrow C + D$
- “Bootstrap compounds”
- Missing initial protein substrates (e.g., ACP)
 - ◆ Protein synthesis not represented

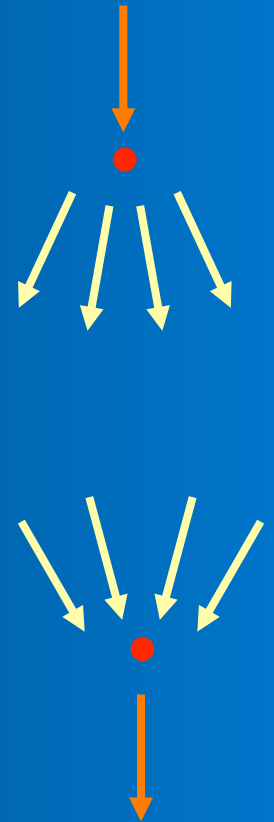
- **Phase III: Forward propagation with 11 more initial metabolites**

- Yielded all 41 essential compounds



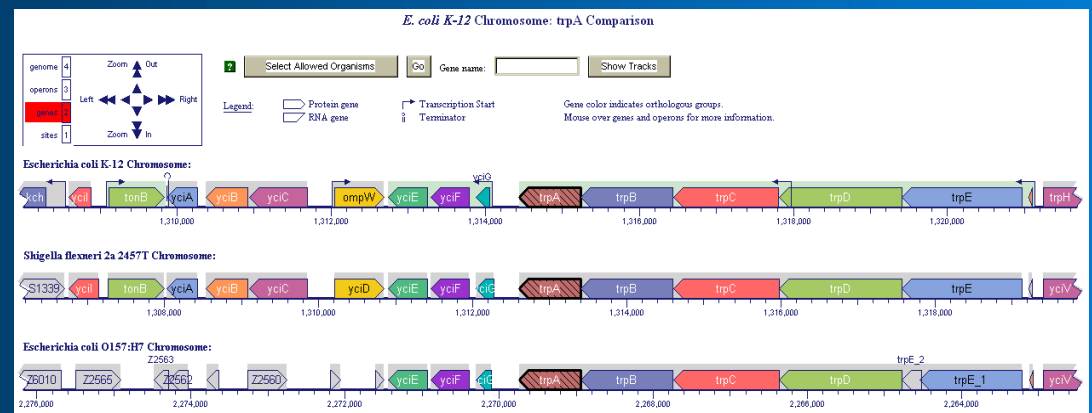
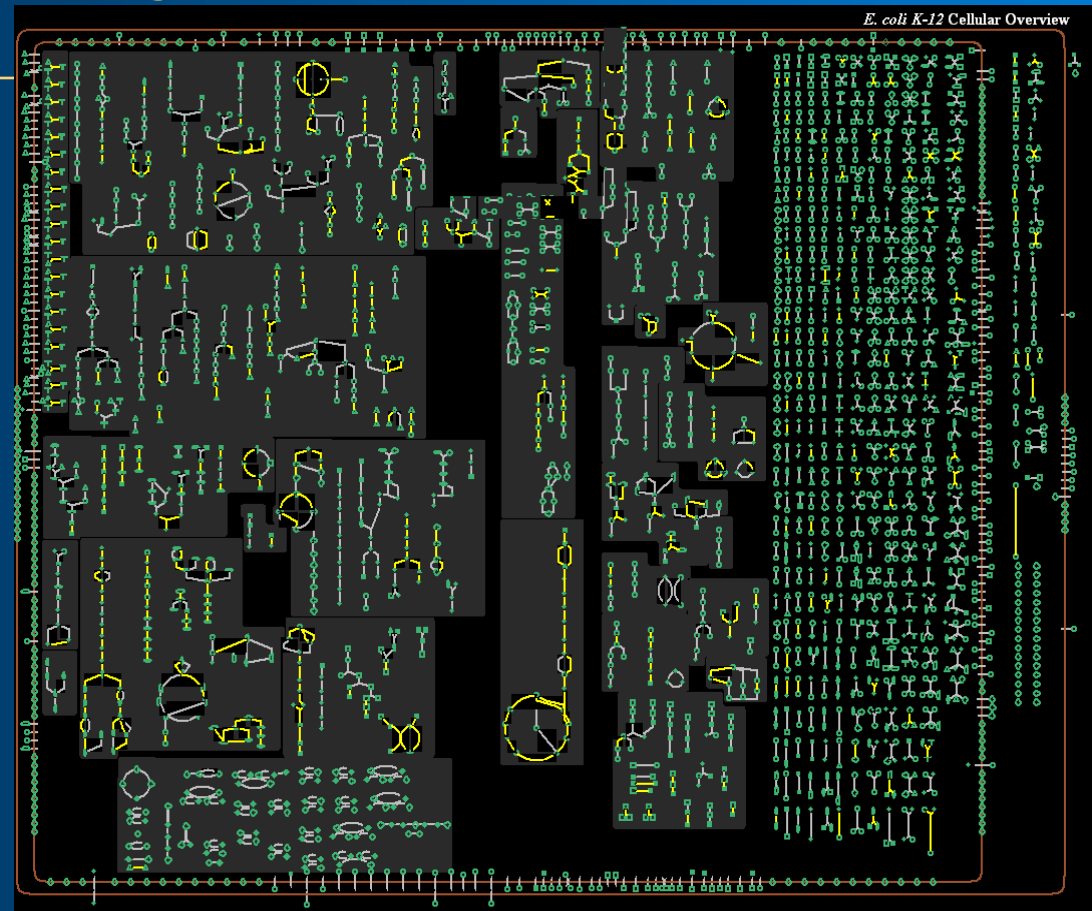
Infer Anti-Microbial Drug Targets

- Infer drug targets as genes coding for enzymes that encode chokepoint reactions
- Two types of **chokepoint** reactions:
- **Chokepoint analysis of *Plasmodium falciparum*:**
 - 216/303 reactions are chokepoints (73%)
 - All 3 clinically proven anti-malarial drugs target chokepoints
 - 21/24 biologically validated drug targets are chokepoints
 - 11.2% of chokepoints are drug targets
 - 3.4% of non-chokepoints are drug targets
 - => Chokepoints are significantly enriched for drug targets



Comparative Analysis

- Via Cellular Overview
- Comparative genome browser
- Comparative pathway table
- Comparative analysis reports
 - Compare reaction complements
 - Compare pathway complements
 - Compare transporter complements



Summary

- **Pathway/Genome Databases**

- MetaCyc non-redundant DB of literature-derived pathways
- Additional organism-specific PGDBs available through SRI at BioCyc.org
- Computational theories of biochemical machinery

- **Pathway Tools software**

- Extract pathways from genomes
- Morph annotated genome into structured ontology
- Distributed curation tools for MODs
- Query, visualization, WWW publishing

How to Learn More

- **BioCyc Webinars**
 - See BioCyc.org
- **BioCyc publications page**
 - BioCyc.org
- **Pathway Tools training course**
- **Pathway Tools feedback sessions**
 - ptools-support@ai.sri.com
- **Try out Pathway Tools**