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

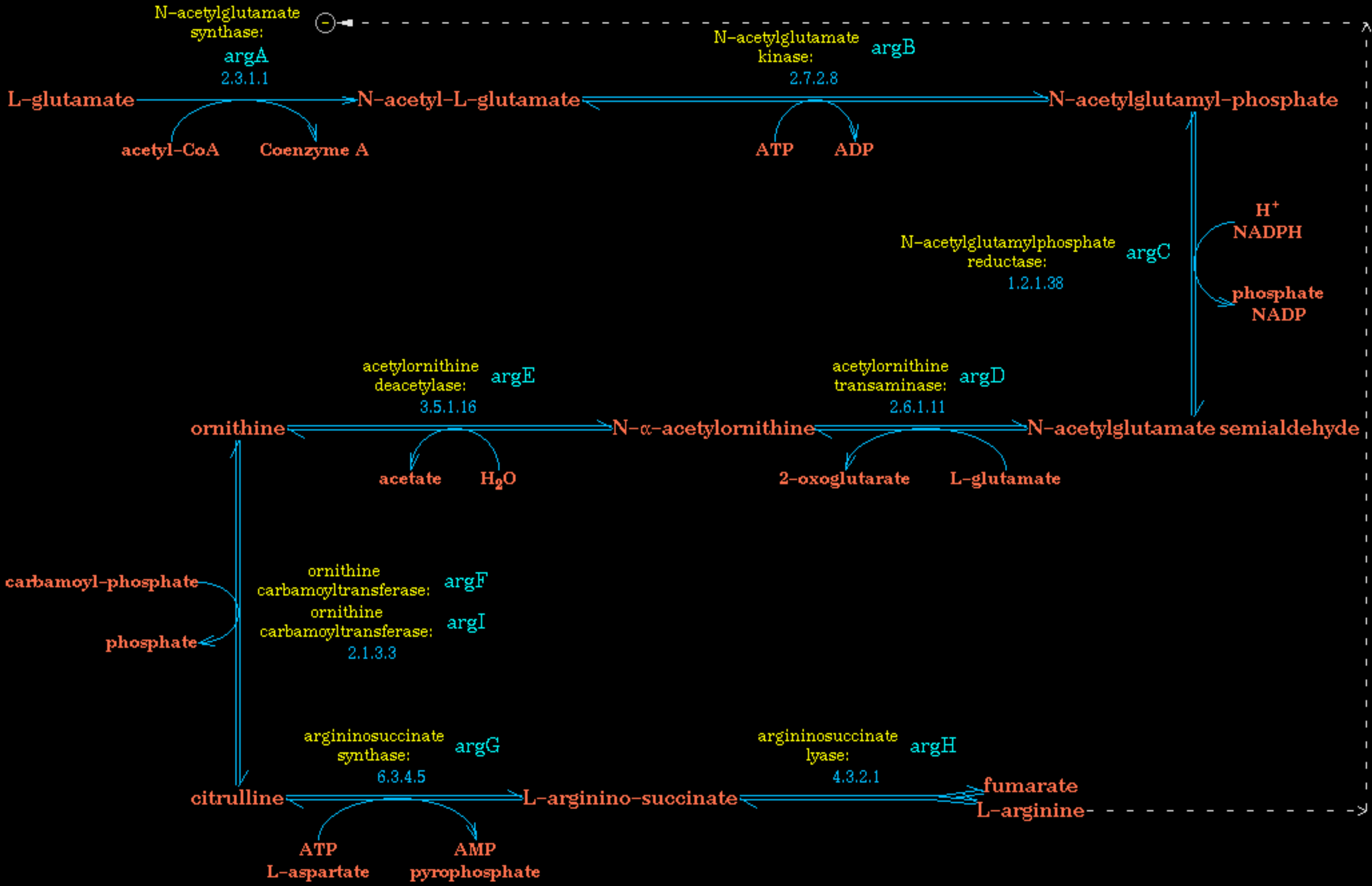
- A chemical reaction interconverts chemical compounds (analogous to a production rule)

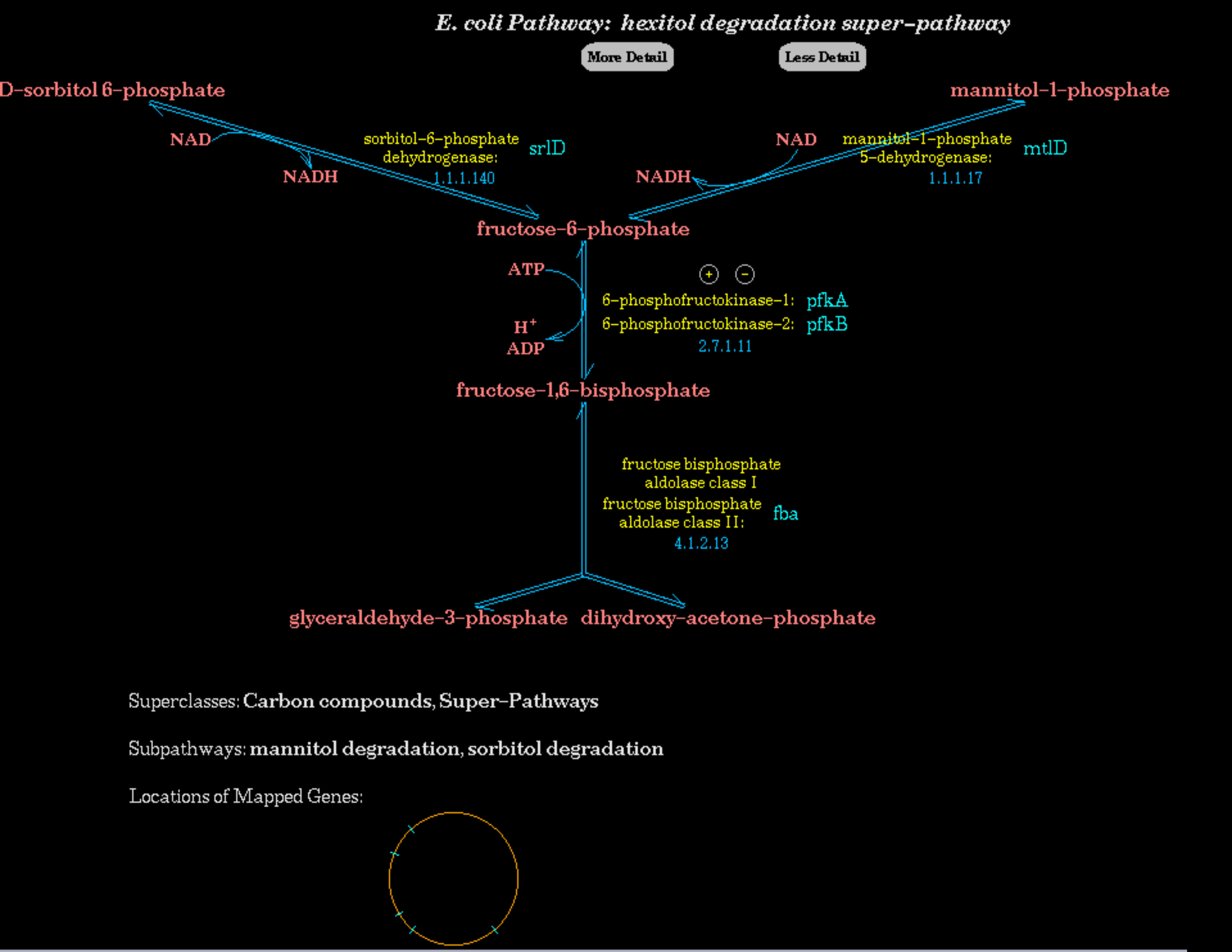
$$A + B = C + D$$

- 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)

$$A \longrightarrow C \longrightarrow E$$


N-acetylglutamate
synthase:
argA
2.3.1.1
L-glutamate
N-acetyl-L-glutamate
acetyl-CoA
Coenzyme A
-
N-acetylglutamate
kinase:
argB
2.7.2.8
N-acetylglutamyl-phosphate
ATP
ADP
N-acetylglutamylphosphate
reductase:
argC
1.2.1.38
H^+
NADPH
phosphate
NADP
acetylornithine
deacetylase:
argE
3.5.1.16
acetylornithine
transaminase:
argD
2.6.1.11
ornithine
N-α-acetylornithine
N-acetylglutamate semialdehyde
acetate
H_2O
2-oxoglutarate
L-glutamate
carbamoyl-phosphate
phosphate
ornithine
carbamoyltransferase:
argF
ornithine
carbamoyltransferase:
argI
2.1.3.3
argininosuccinate
synthase:
argG
6.3.4.5
argininosuccinate
lyase:
argH
4.3.2.1
citrulline
L-arginino-succinate
fumarate
L-arginine
ATP
L-aspartate
AMP
pyrophosphate



E. coli Pathway: hexitol degradation super-pathway

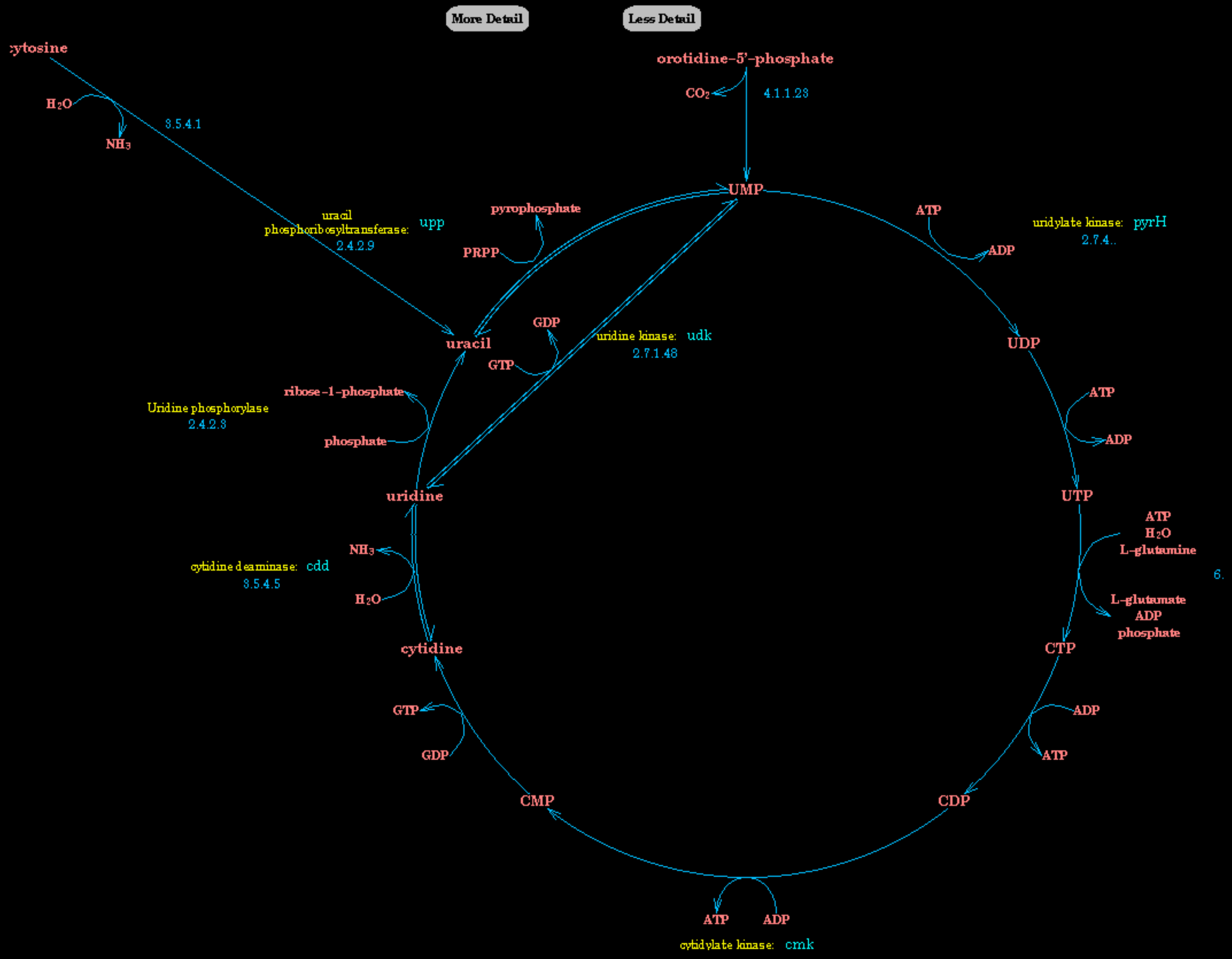
More Detail
Less Detail
D-sorbitol 6-phosphate
NAD
NADH
sorbitol-6-phosphate dehydrogenase:
srlD
1.1.1.140
mannitol-1-phosphate
NAD
NADH
mannitol-1-phosphate 5-dehydrogenase:
mtlD
1.1.1.17
fructose-6-phosphate
ATP
H+
ADP
+
-
6-phosphofructokinase-1:
pfkA
6-phosphofructokinase-2:
pfkB
2.7.1.11
fructose-1,6-bisphosphate
fructose bisphosphate aldolase class I
fructose bisphosphate aldolase class II:
fba
4.1.2.13
glyceraldehyde-3-phosphate
dihydroxy-acetone-phosphate

Superclasses: **Carbon compounds, Super-Pathways**

Subpathways: **mannitol degradation, sorbitol degradation**

Locations of Mapped Genes:

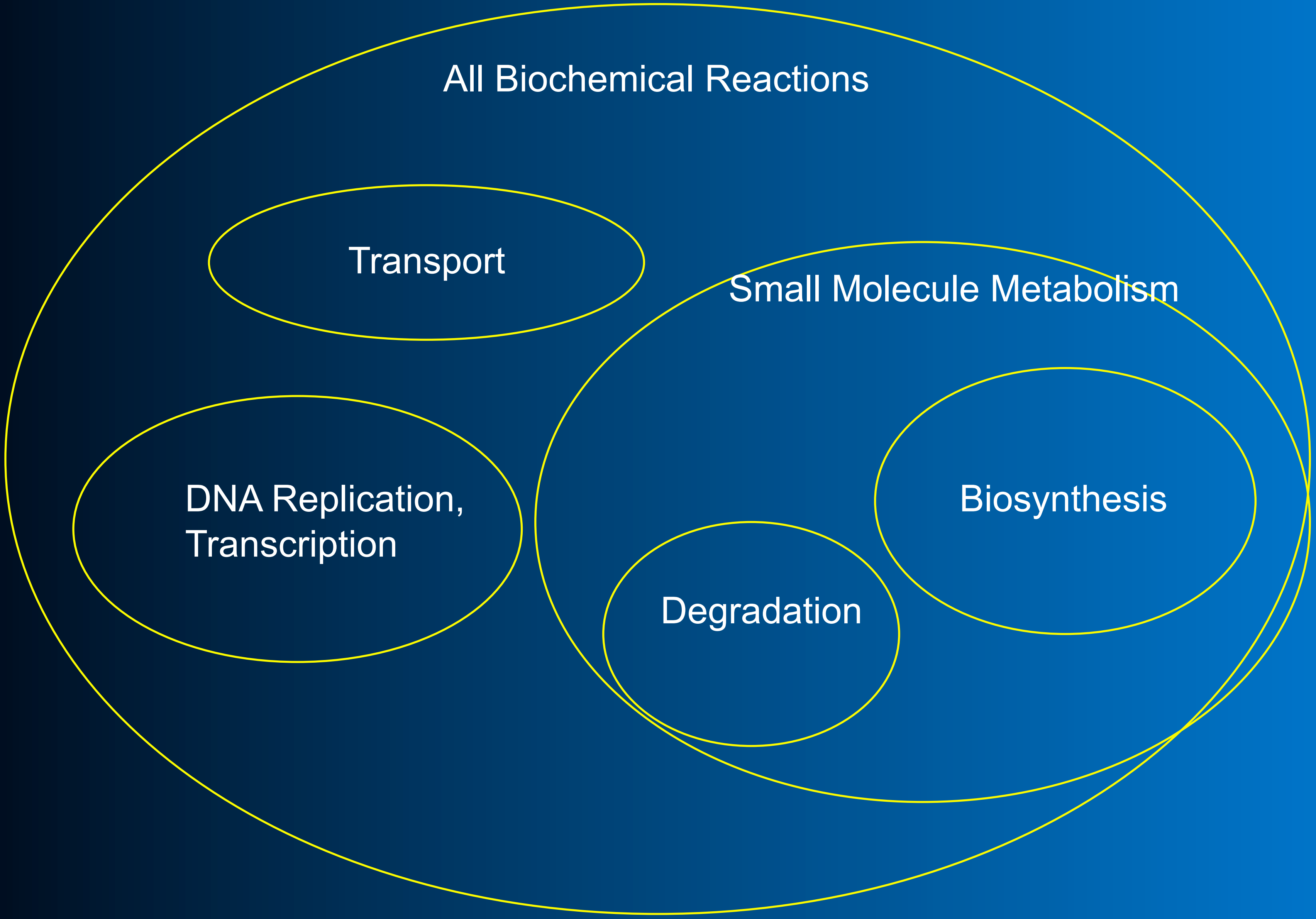
Mp. pneumoniae Pathway: pyrimidine ribonucleotide/ribonucleoside metabolism



Definition of Small-Molecule Metabolism

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

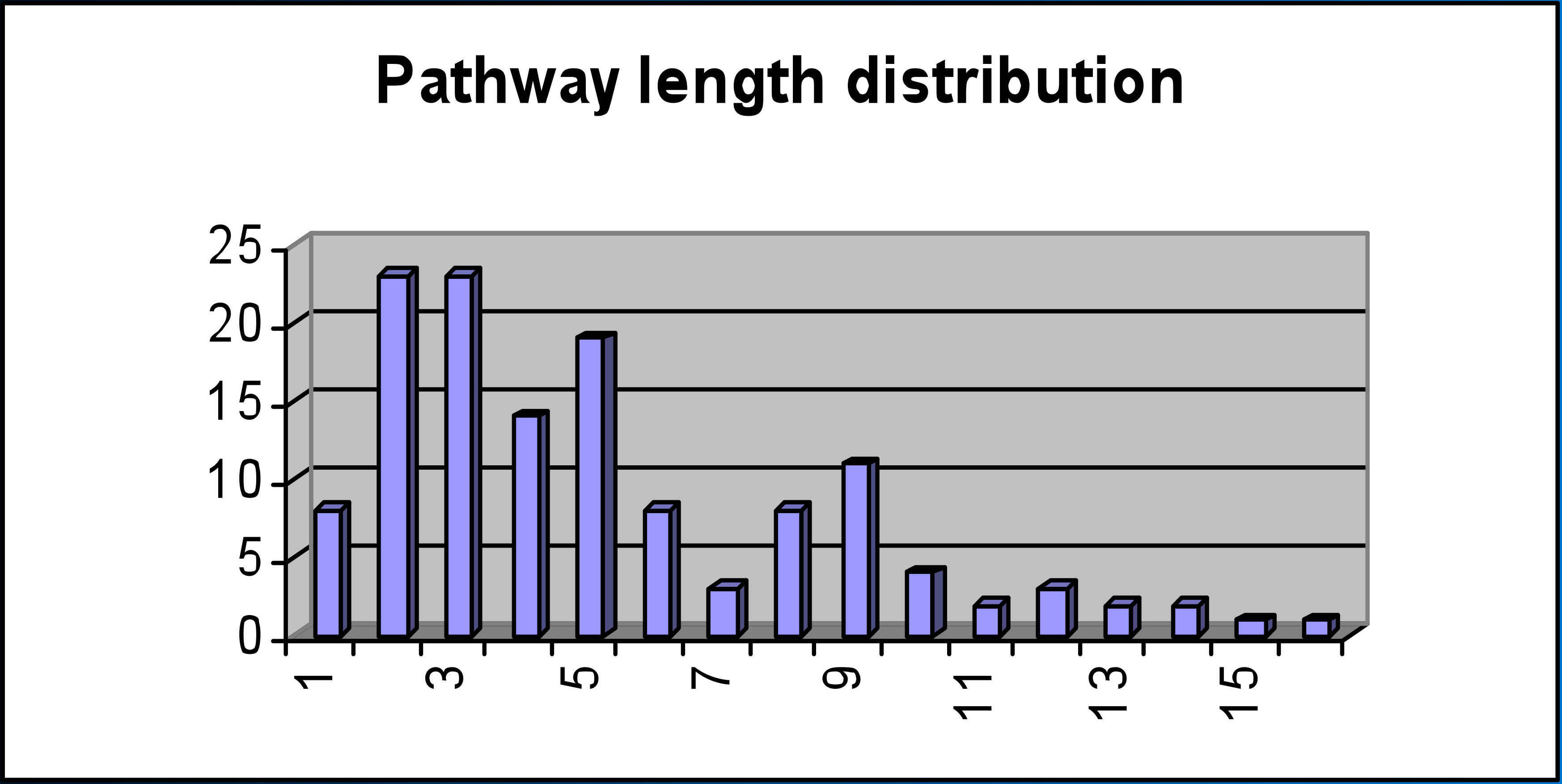
Small Molecule Metabolism



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

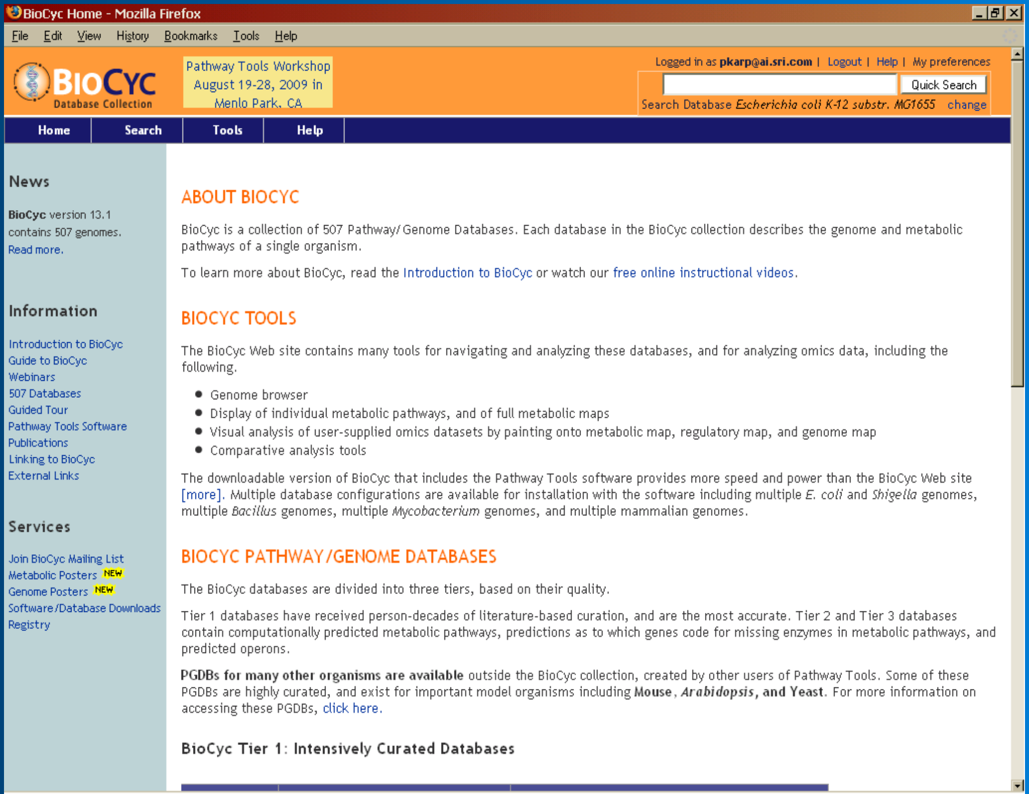
BioCyc Collection of 507 Pathway/ Genome Databases

- **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**



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

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
Genome Posters
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

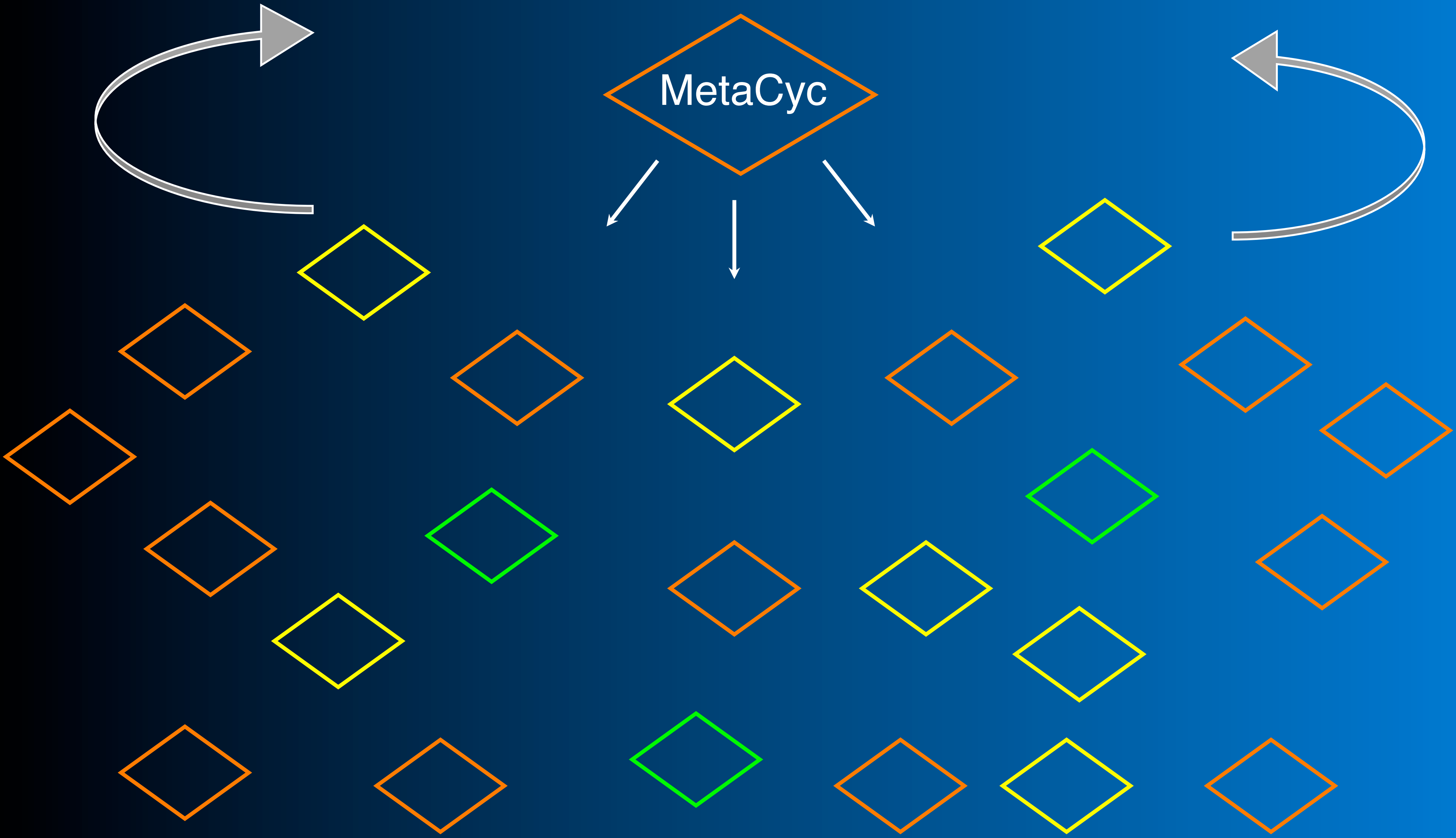
The BioCyc databases are divided into three tiers, based on their quality.

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.

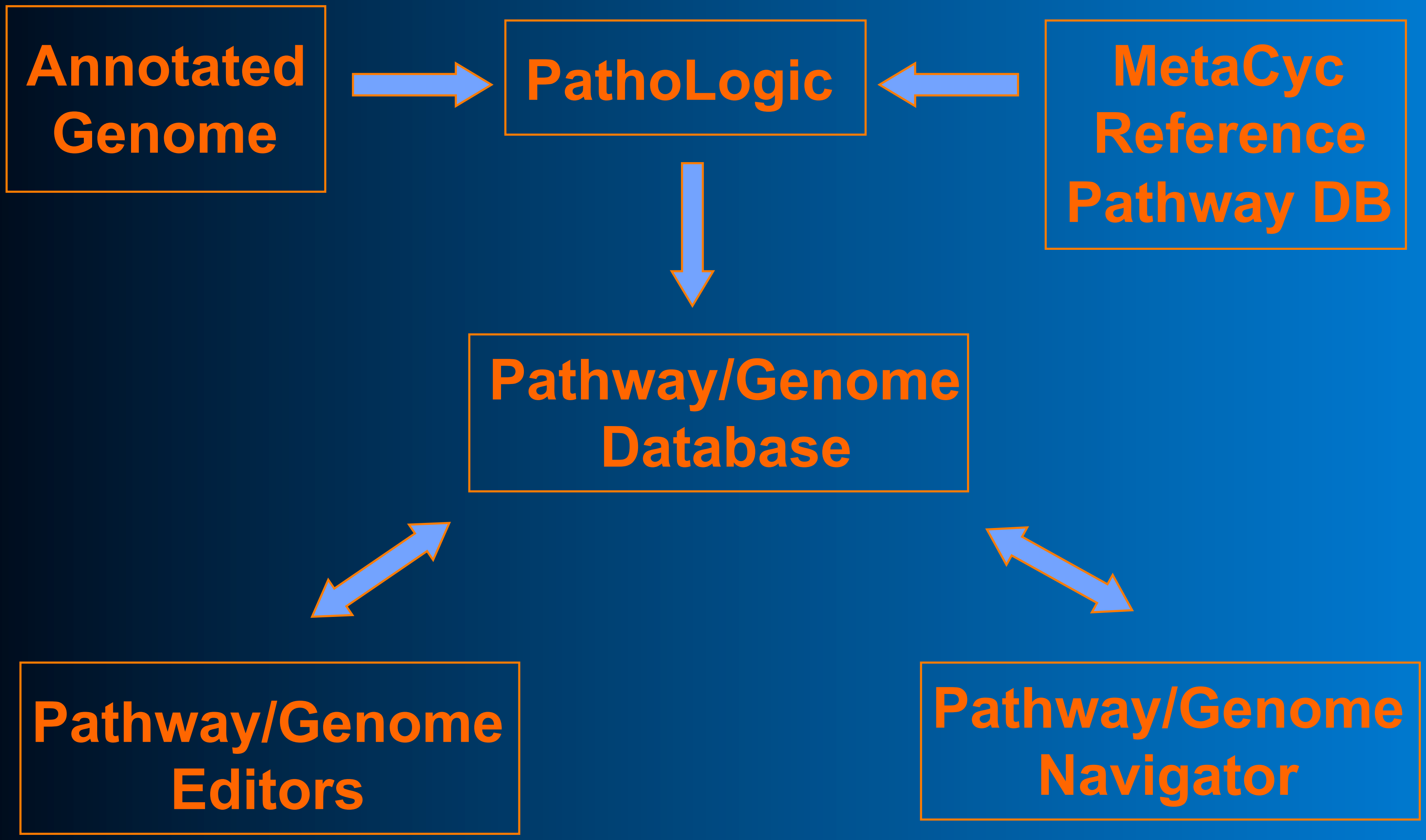
PGDBs for many other organisms are available outside the BioCyc collection, created by other users of Pathway Tools. Some of these PGDBs are highly curated, and exist for important model organisms including **Mouse**, ***Arabidopsis***, **and Yeast**. For more information on accessing these PGDBs, click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

BioCyc Tier 1: Intensively Curated Databases

Family of Pathway/Genome Databases



Pathway Tools Overview



Annotated Genome

PathoLogic

MetaCyc Reference Pathway DB

Pathway/Genome Database

Pathway/Genome Editors

Pathway/Genome Navigator

Briefings in Bioinformatics 11:40-79 2010

Pathway Tools Software: PathoLogic

- **Computational creation of new Pathway/Genome Databases**

- **Transforms genome into Pathway Tools schema and layers inferred information above the genome**

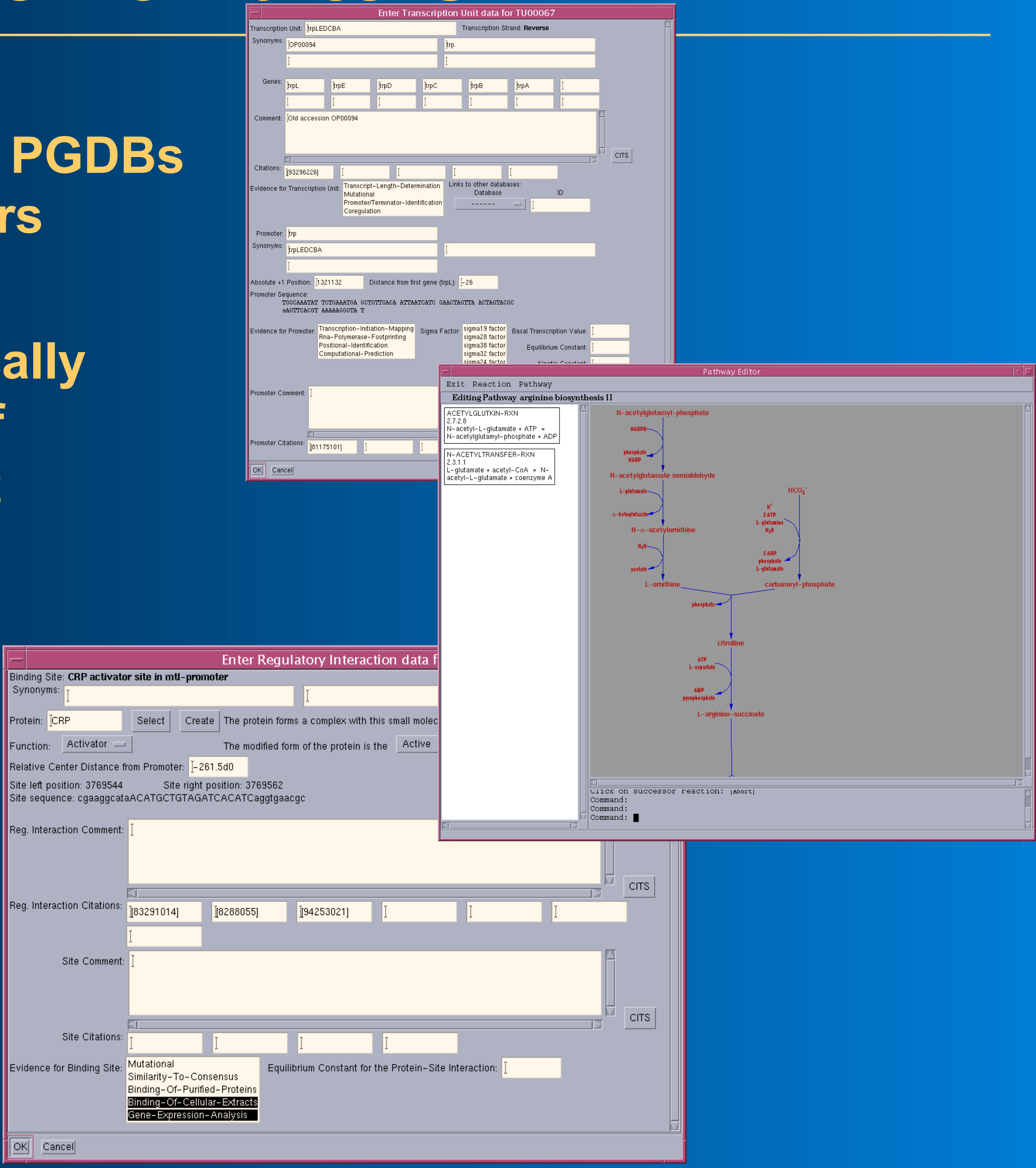
- **Predicts operons**
- **Predicts metabolic network**
- **Predicts pathway hole fillers**
- **Infers transport reactions**

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



Enter Transcription Unit data for TU00067

Transcription Unit: trpLEDCBA Transcription Strand: **Reverse**

Synonyms: OP00094 trp

Genes: trpL trpE trpD trpC trpB trpA

Comment: Old accession OP00094

Citations: [93296226]

Evidence for Transcription Unit: Transcript-Length-Determination
Mutational
Promoter/Terminator-Identification
Coregulation

Links to other databases:
Database ------ ID

Promoter: trp

Synonyms: trpLEDCBA

Absolute +1 Position: 1321132 Distance from first gene (trpL): -26

Promoter Sequence:

```
TGGCAAATAT TCTGAAATGA GCTGTTGACA ATTAATCATC GAACTAGTTA ACTAGTACGC
aAGTTCACGT AAAAAGGGTA T
```

Evidence for Promoter: Transcription-Initiation-Mapping
Rna-Polymerase-Footprinting
Positional-Identification
Computational-Prediction

Sigma Factor: sigma19 factor
sigma28 factor
sigma38 factor
sigma32 factor
sigma24 factor

Basal Transcription Value:

Equilibrium Constant:

Kinetic Constant:

Promoter Comment:

Promoter Citations: [81175101]

Pathway Editor

Editing Pathway arginine biosynthesis II

ACETYLGLUTKIN-RXN
2.7.2.8
N-acetyl-L-glutamate + ATP = N-acetylglutamyl-phosphate + ADP

N-ACETYLTRANSFER-RXN
2.3.1.1
L-glutamate + acetyl-CoA = N-acetyl-L-glutamate + coenzyme A

```
Click on successor reaction: [Abort]
Command:
Command:
Command:
```

Enter Regulatory Interaction data f

Binding Site: **CRP activator site in mtl-promoter**

Synonyms:

Protein: CRP Select Create The protein forms a complex with this small molec

Function: Activator The modified form of the protein is the Active

Relative Center Distance from Promoter: -261.5d0

Site left position: 3769544 Site right position: 3769562

Site sequence: cgaaggcataACATGCTGTAGATCACATCaggtgaacgc

Reg. Interaction Comment:

Reg. Interaction Citations: [83291014] [8288055] [94253021]

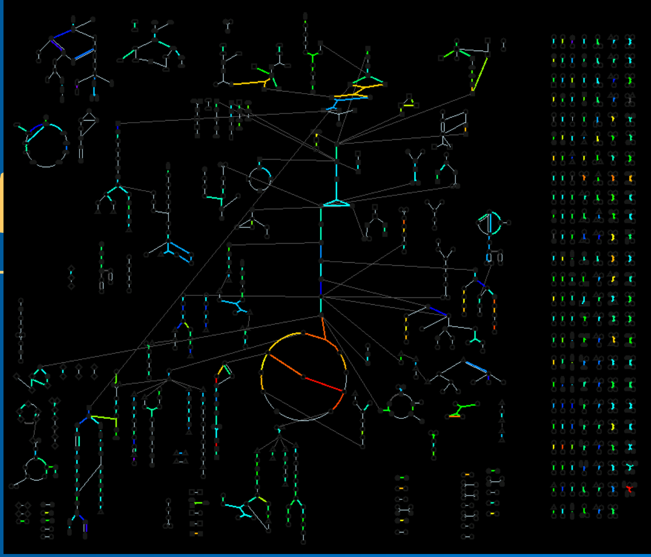
Site Comment:

Site Citations:

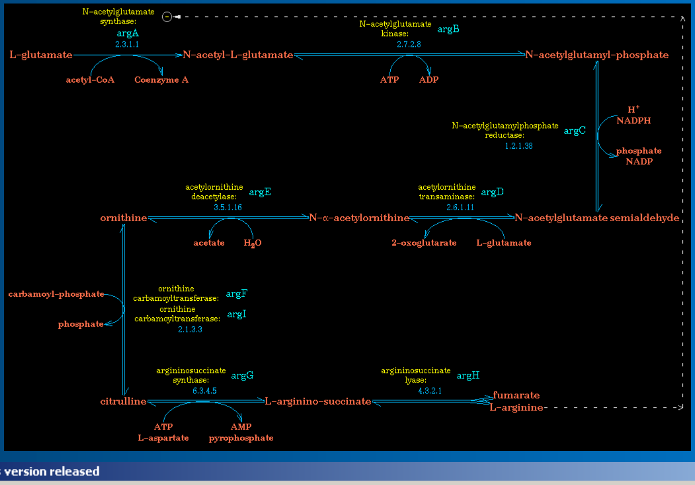
Evidence for Binding Site: Mutational
Similarity-To-Consensus
Binding-Of-Purified-Proteins
Binding-Of-Cellular-Extracts
Gene-Expression-Analysis

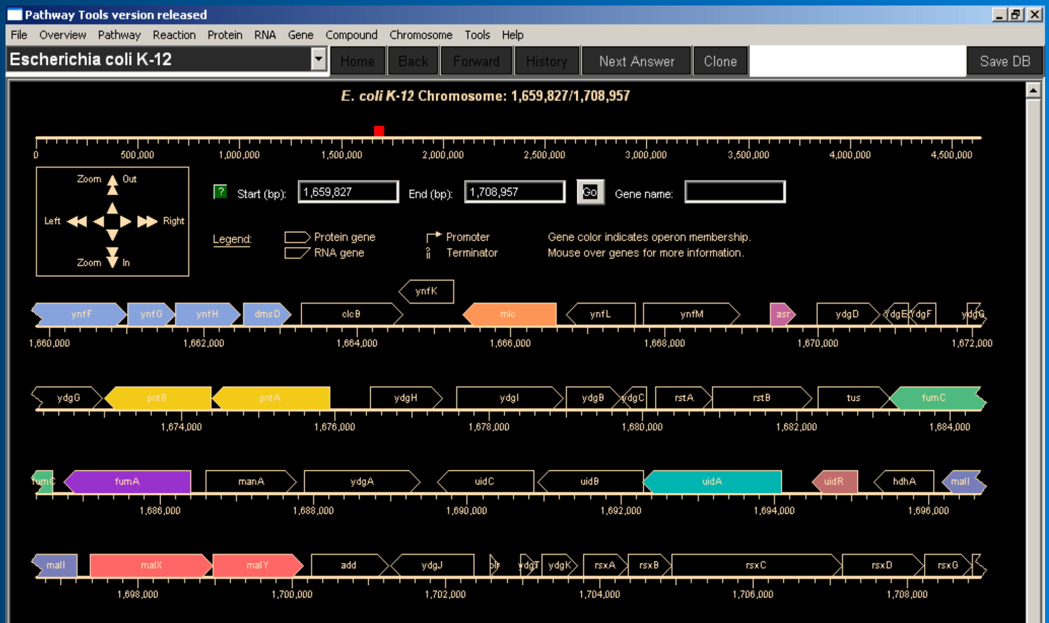
Equilibrium Constant for the Protein-Site Interaction:

Pathway Tools Software: Pathway/Genome Navigat



- **Querying, visualization of pathways, chromosomes, operons**
- **Analysis operations**
  - Pathway visualization of gene-expression data
  - Global comparisons of metabolic networks
  - Comparative genomics
- **WWW publishing of PGDBs**
- **Desktop operation**




Pathway Tools version released

Escherichia coli K-12

E. coli K-12 Chromosome: 1,659,827/1,708,957

Start (bp): 1,659,827 End (bp): 1,708,957 Gene name:

Legend: Protein gene RNA gene Promoter Terminator

Gene color indicates operon membership.
Mouse over genes for more information.

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

Nucleic Acids Research 34:D511-D516 2006

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

Nuc. Acids. Res. **33:D334 2005** ***ASM News*** **70:25 2004** ***Science*** **293:2040**

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

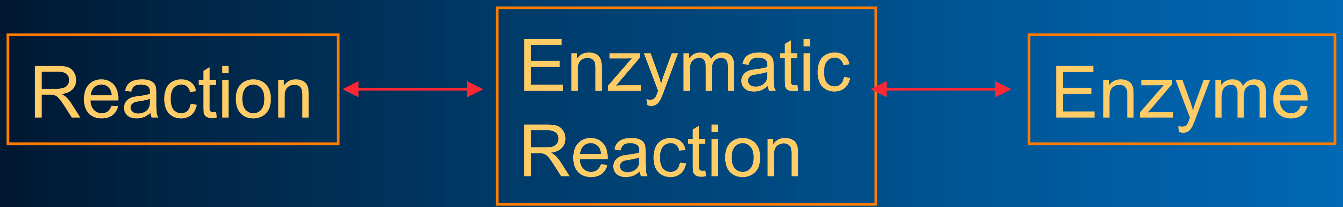
Pathway Tools Schema and Semantic Inference Layer

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**



Reaction

Enzymatic Reaction

Enzyme

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

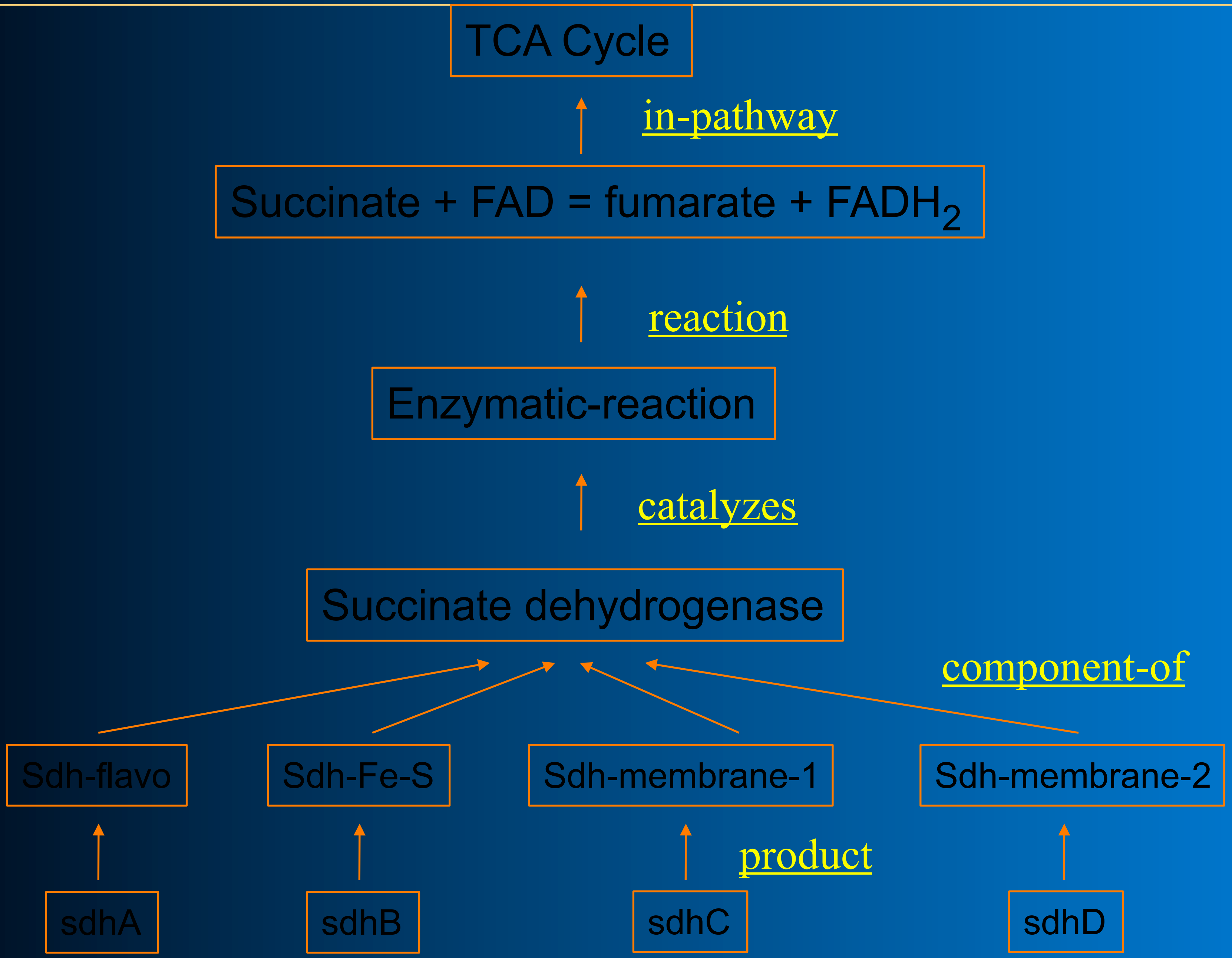
- **Class names are usually capitalized, plural, separated by dashes**

- **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



TCA Cycle

in-pathway

Succinate + FAD = fumarate + $FADH_2$

reaction

Enzymatic-reaction

catalyzes

Succinate dehydrogenase

component-of

Sdh-flavo

Sdh-Fe-S

Sdh-membrane-1

Sdh-membrane-2

product

sdhA

sdhB

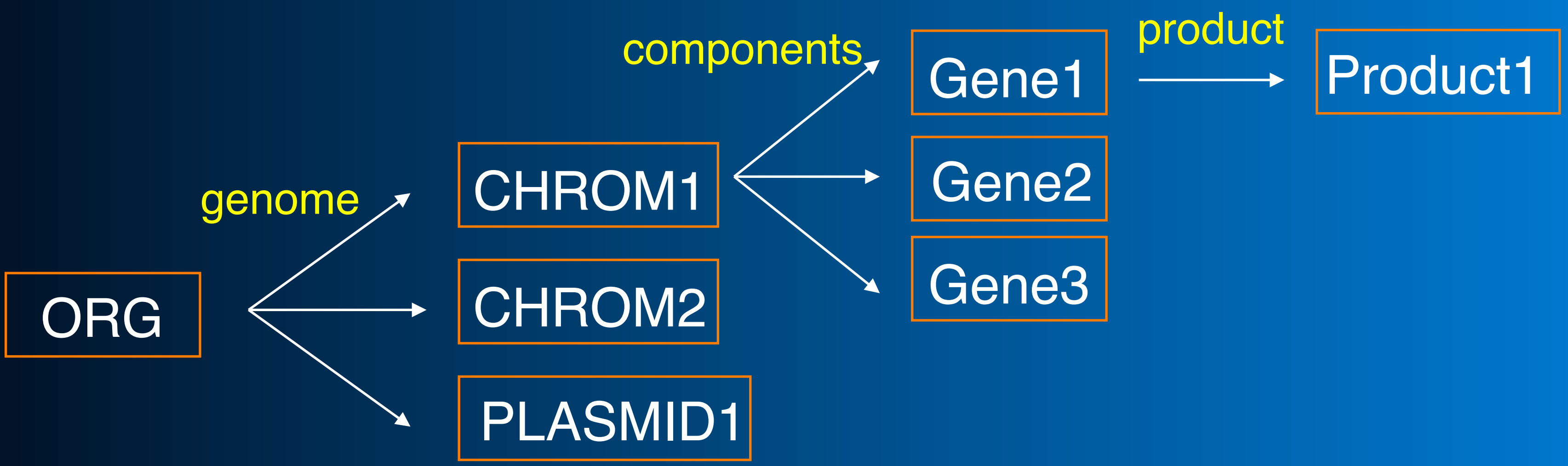
sdhC

sdhD

Pathway Tools Schema and Semantic Inference Layer

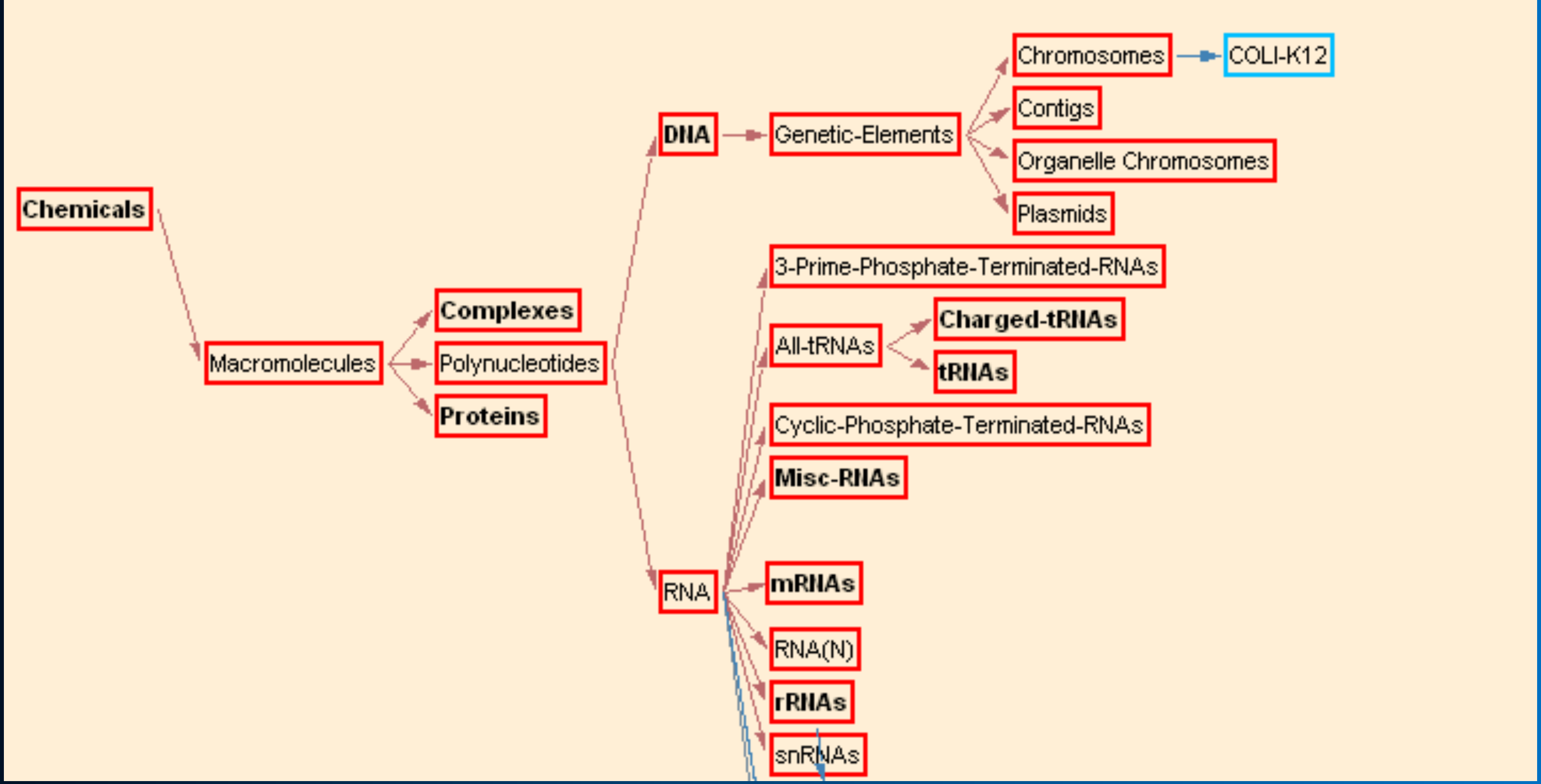
Genes, Operons, and Replicons

Representing a Genome



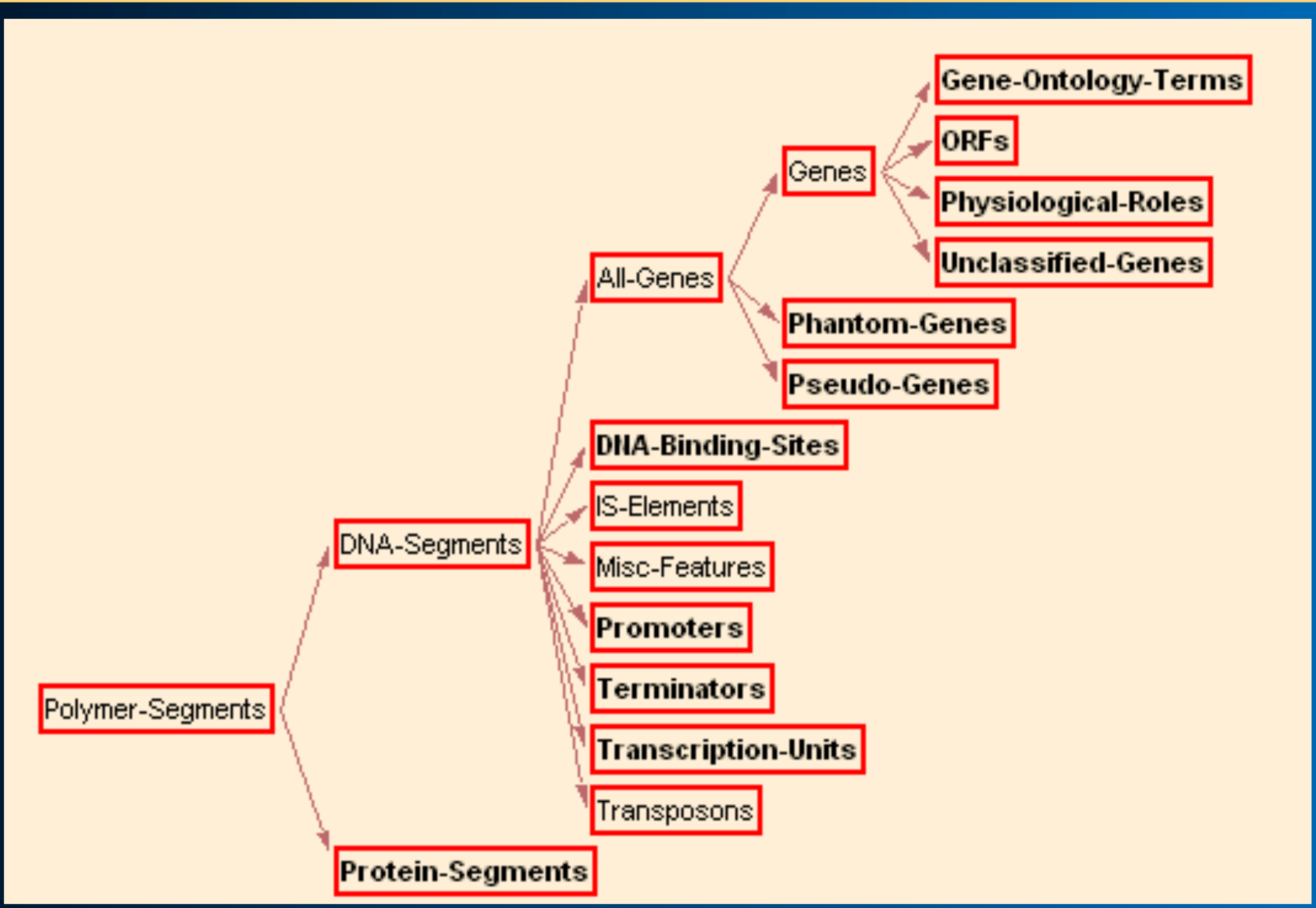
- **Classes:**
  - ORG is of class Organisms
  - CHROM1 is of class Chromosomes
  - PLASMID1 is of class Plasmids
  - Gene1 is of class Genes
  - Product1 is of class Polypeptides or RNA

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)**

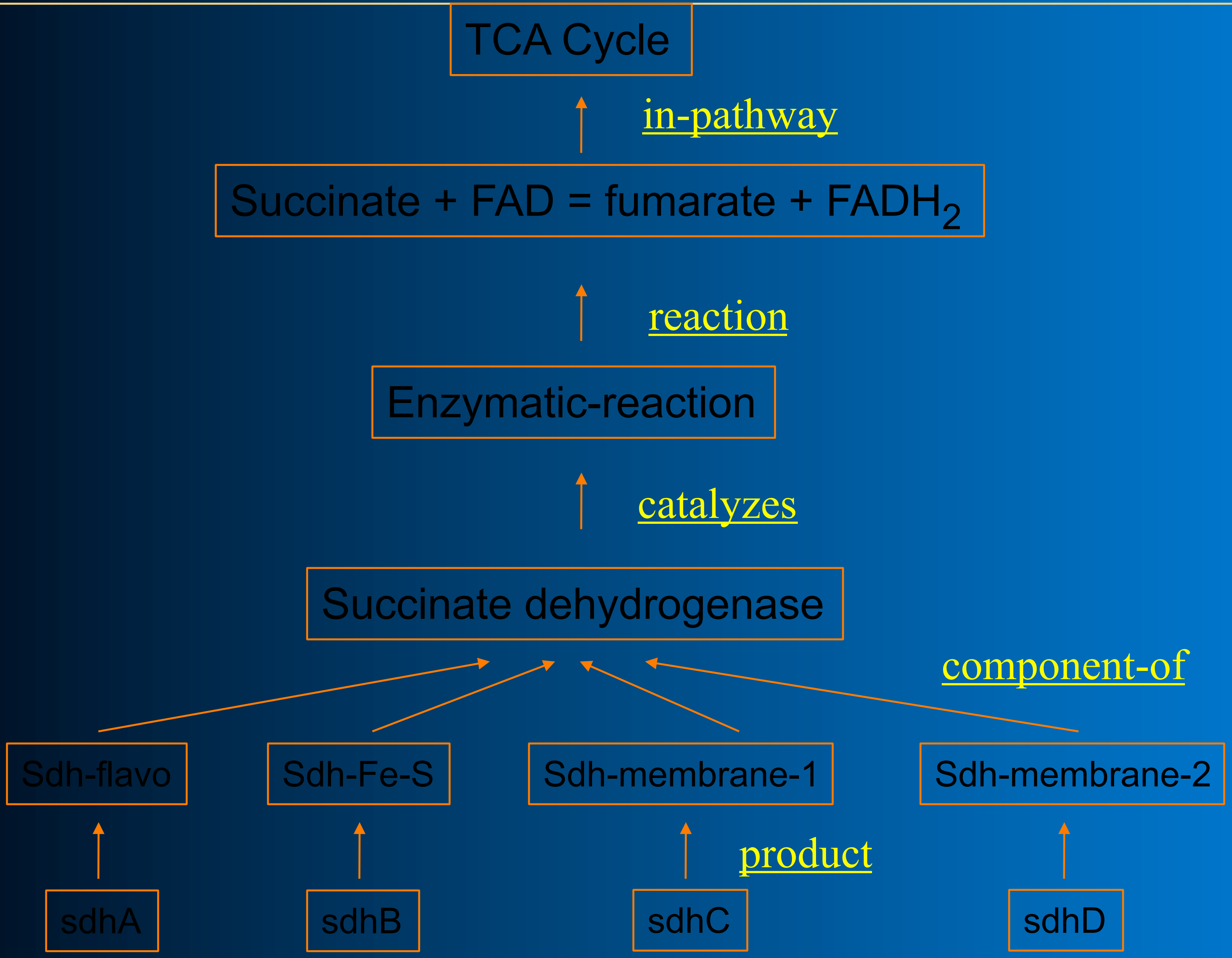
- **Protein complex: proteins consisting of multiple polypeptides or protein complexes**

- **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**
- **pI, 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 enzrxns 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

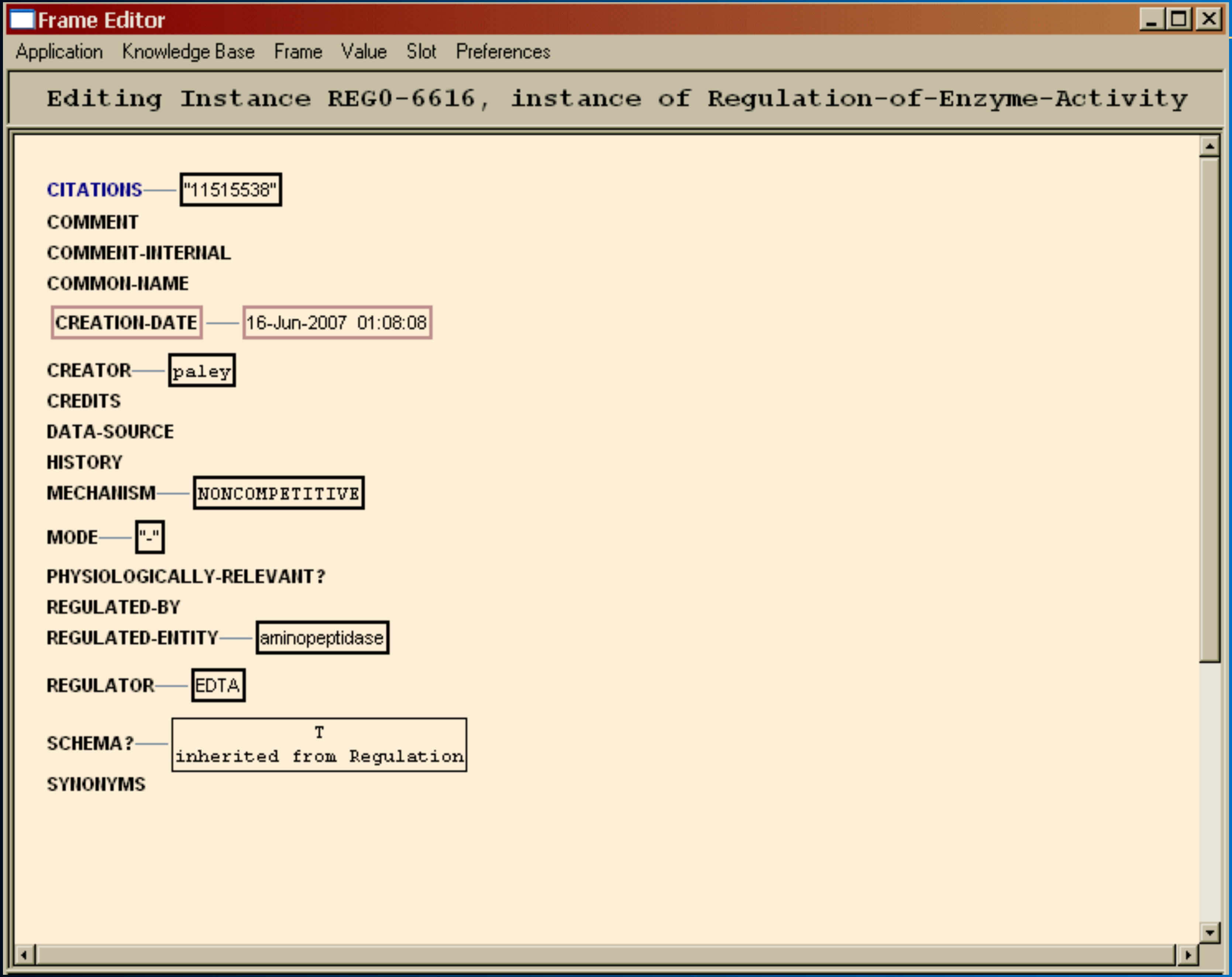
Semantic Inference Layer

- **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)**

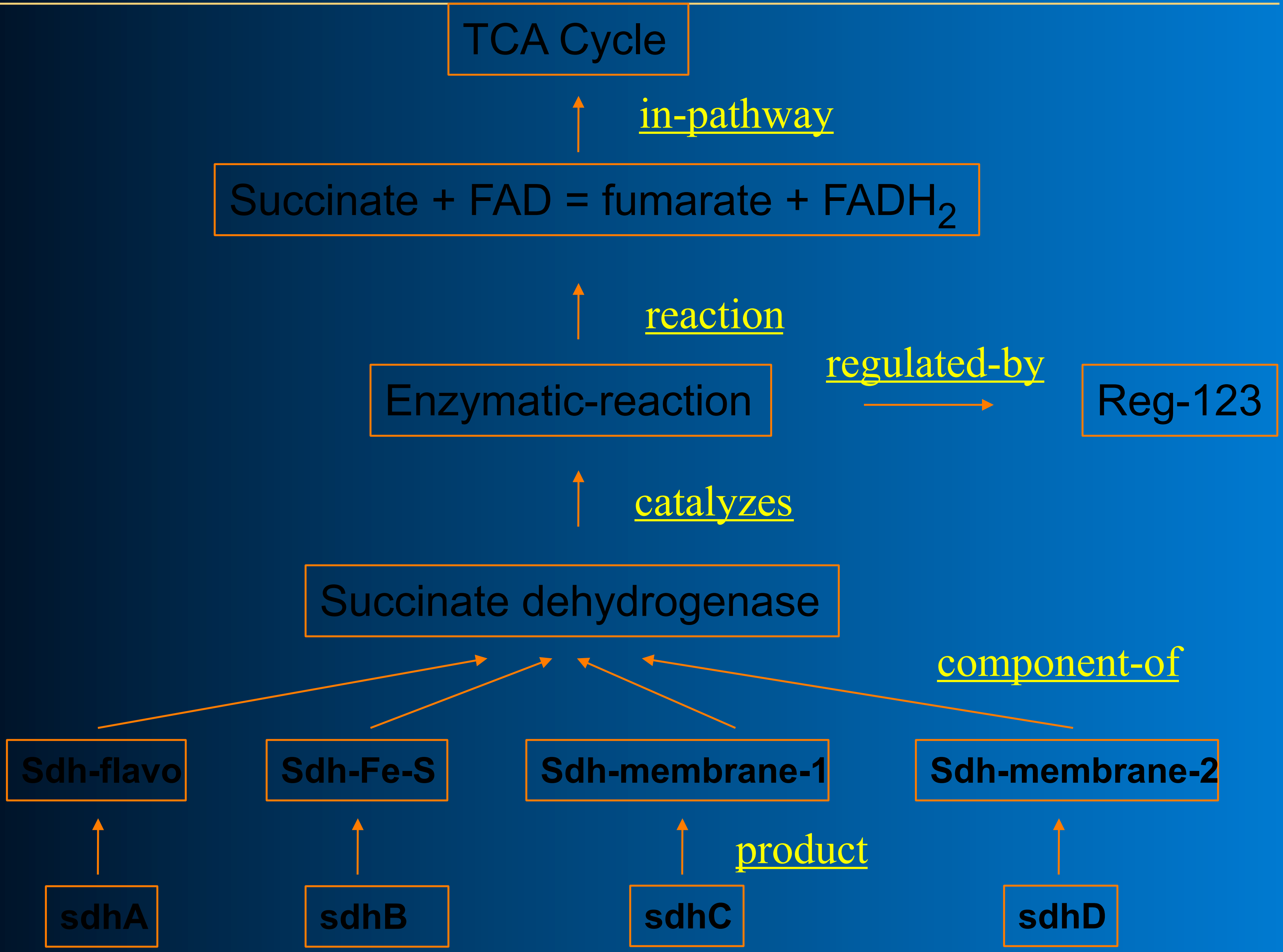
Regulation of Enzyme Activity



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



TCA Cycle

in-pathway

Succinate + FAD = fumarate + $FADH_2$

reaction

Enzymatic-reaction

regulated-by

Reg-123

catalyzes

Succinate dehydrogenase

component-of

Sdh-flavo

Sdh-Fe-S

Sdh-membrane-1

Sdh-membrane-2

product

sdhA

sdhB

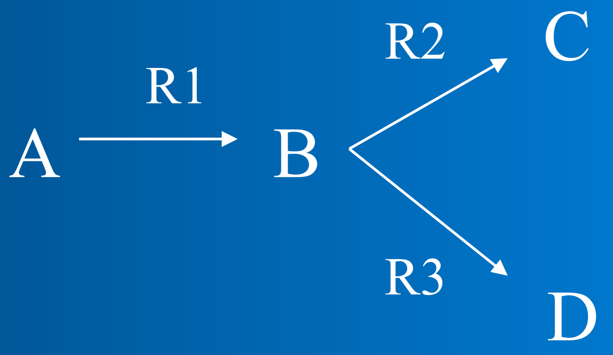
sdhC

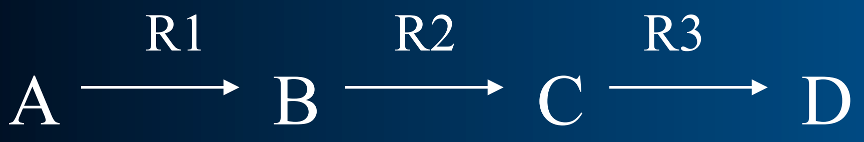
sdhD

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)

ISMB-94, Bioinformatics 16:269 2000

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

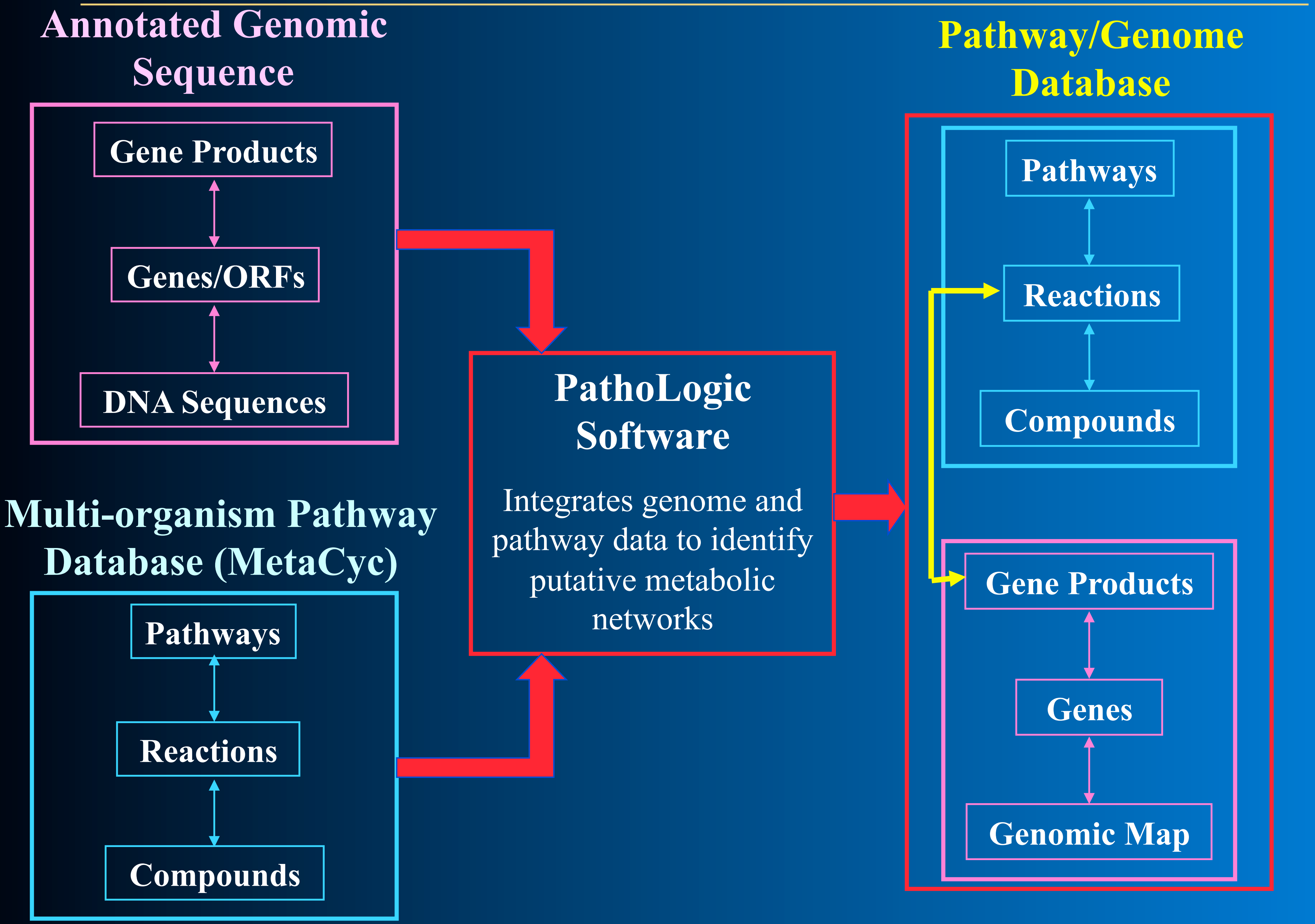
Inference Algorithms

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



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

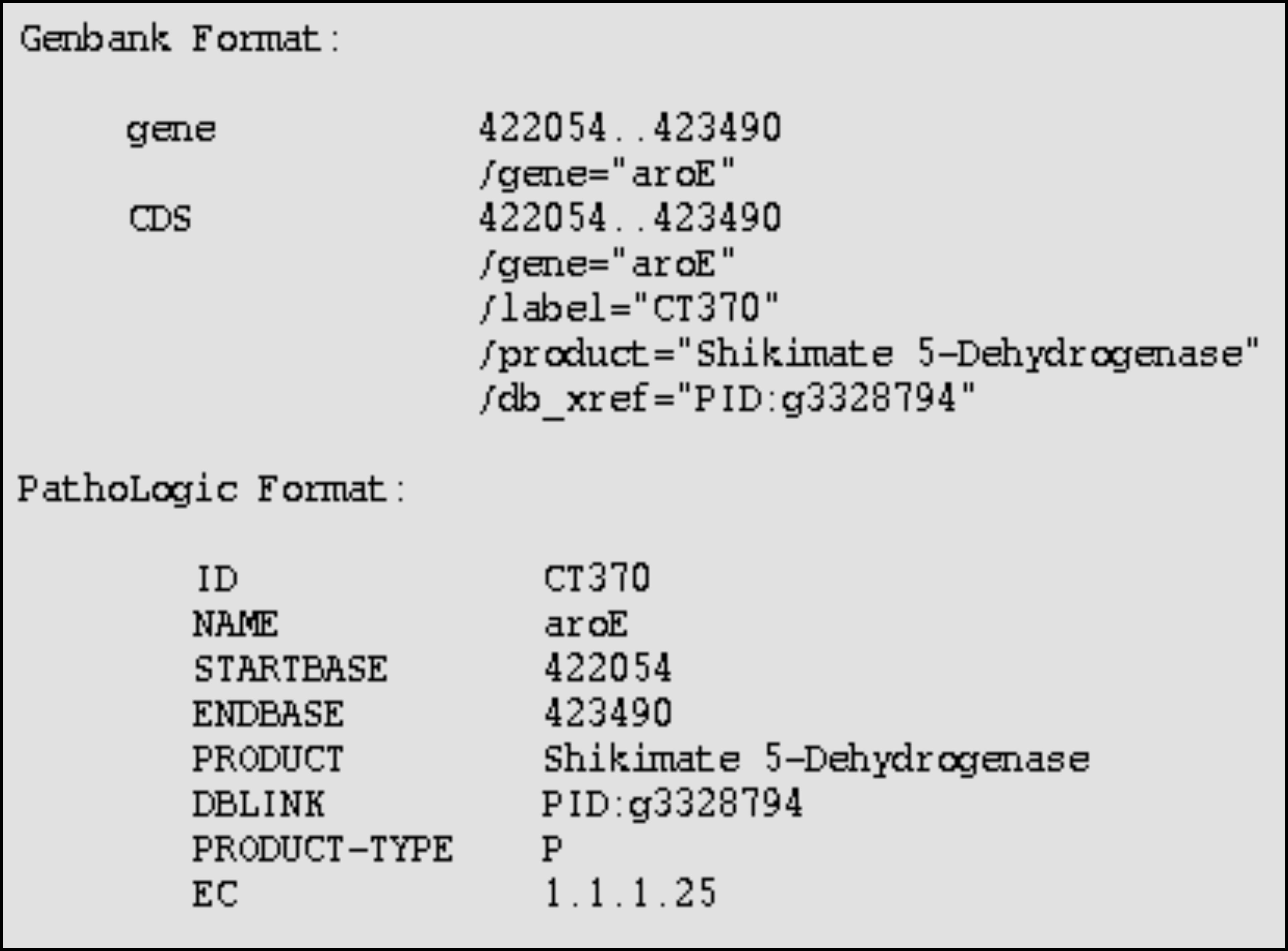
Reactions

Compounds

Gene Products

Genes

Genomic Map



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
```

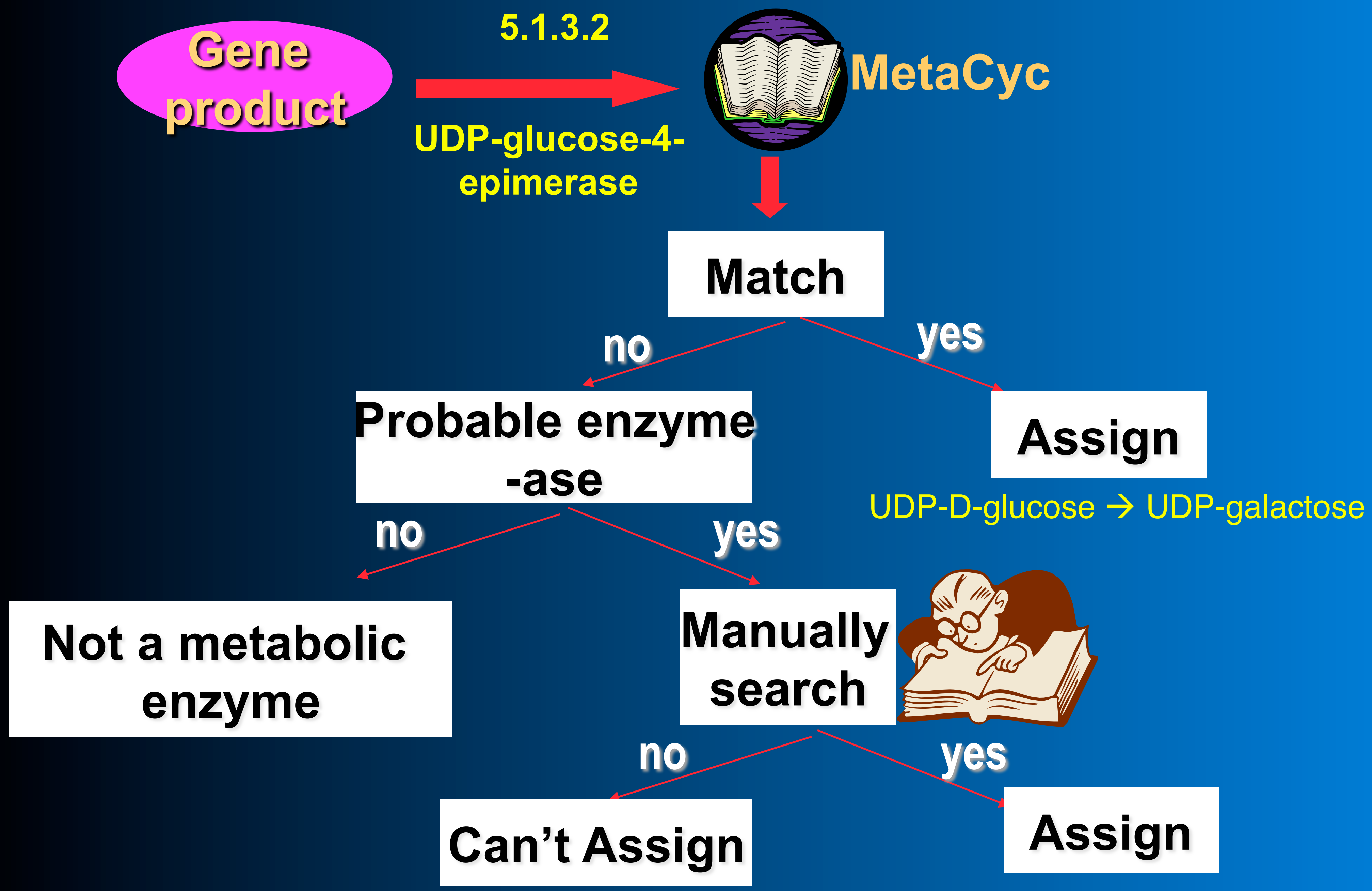
Pathway Prediction

- 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



Gene product

5.1.3.2

UDP-glucose-4-epimerase

MetaCyc

Match

no

yes

Probable enzyme -ase

Assign

UDP-D-glucose → UDP-galactose

no

yes

Not a metabolic enzyme

Manually search

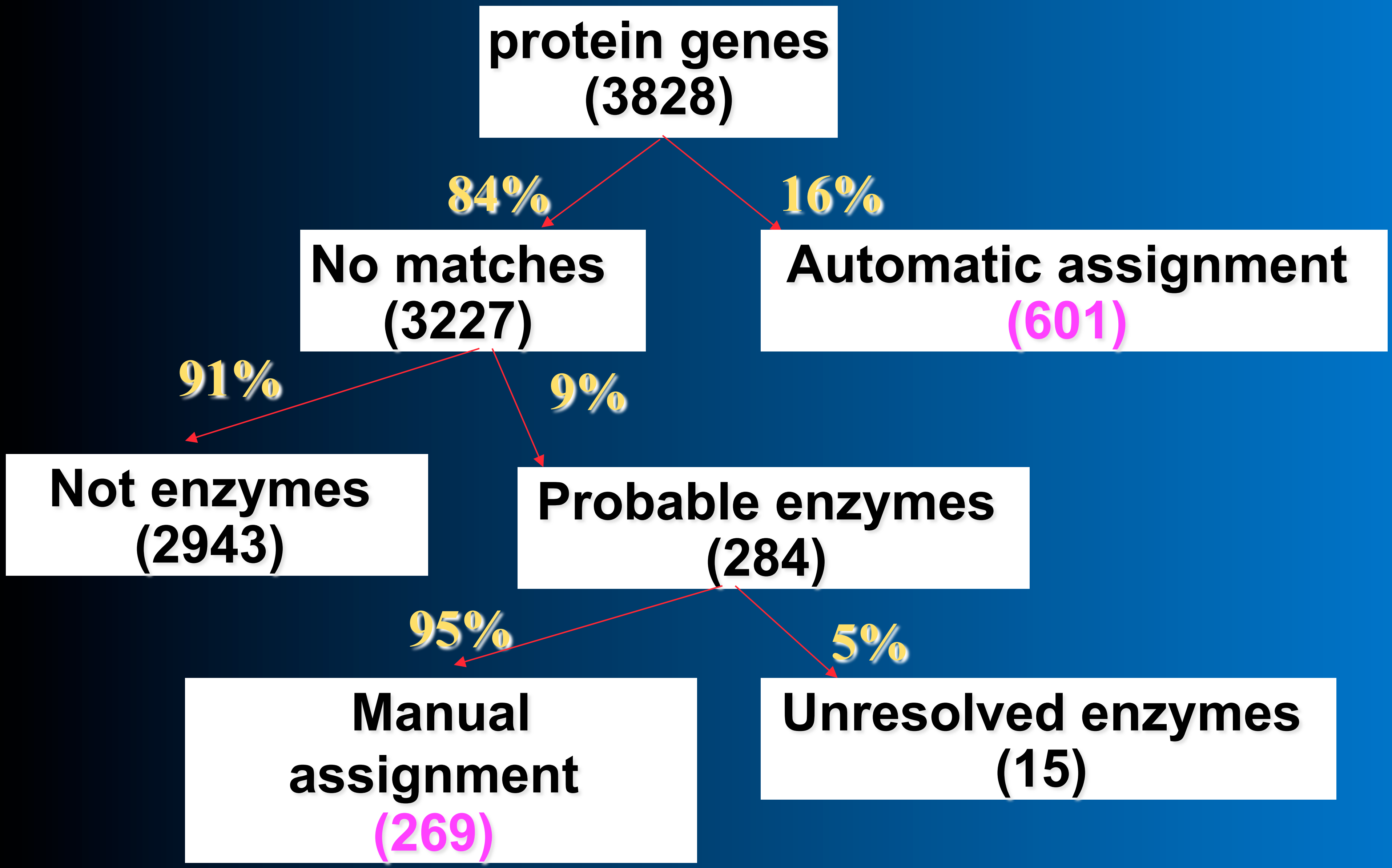
no

yes

Can’t Assign

Assign

Vibrio cholerae Enzyme Matching Results



protein genes (3828)

84%

16%

No matches (3227)

Automatic assignment (601)

91%

9%

Not enzymes (2943)

Probable enzymes (284)

95%

5%

Manual assignment (269)

Unresolved enzymes (15)

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

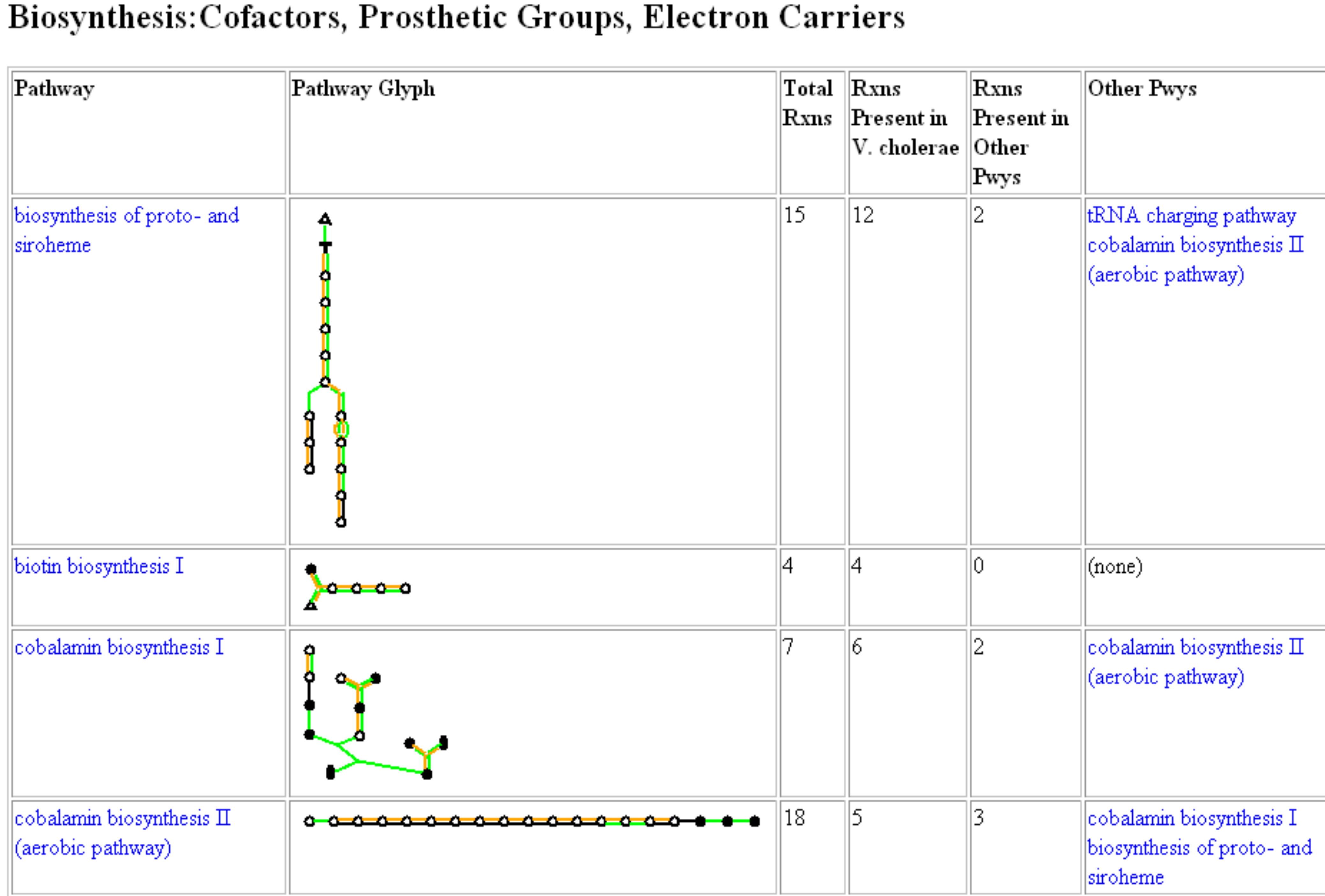
- **For each P in U:**
  - If current organism is outside taxonomic range of P AND at least one reaction in P lacks an enzyme, delete P from U
  - If all reactions of P designated as key reactions have no enzyme, delete P from U

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

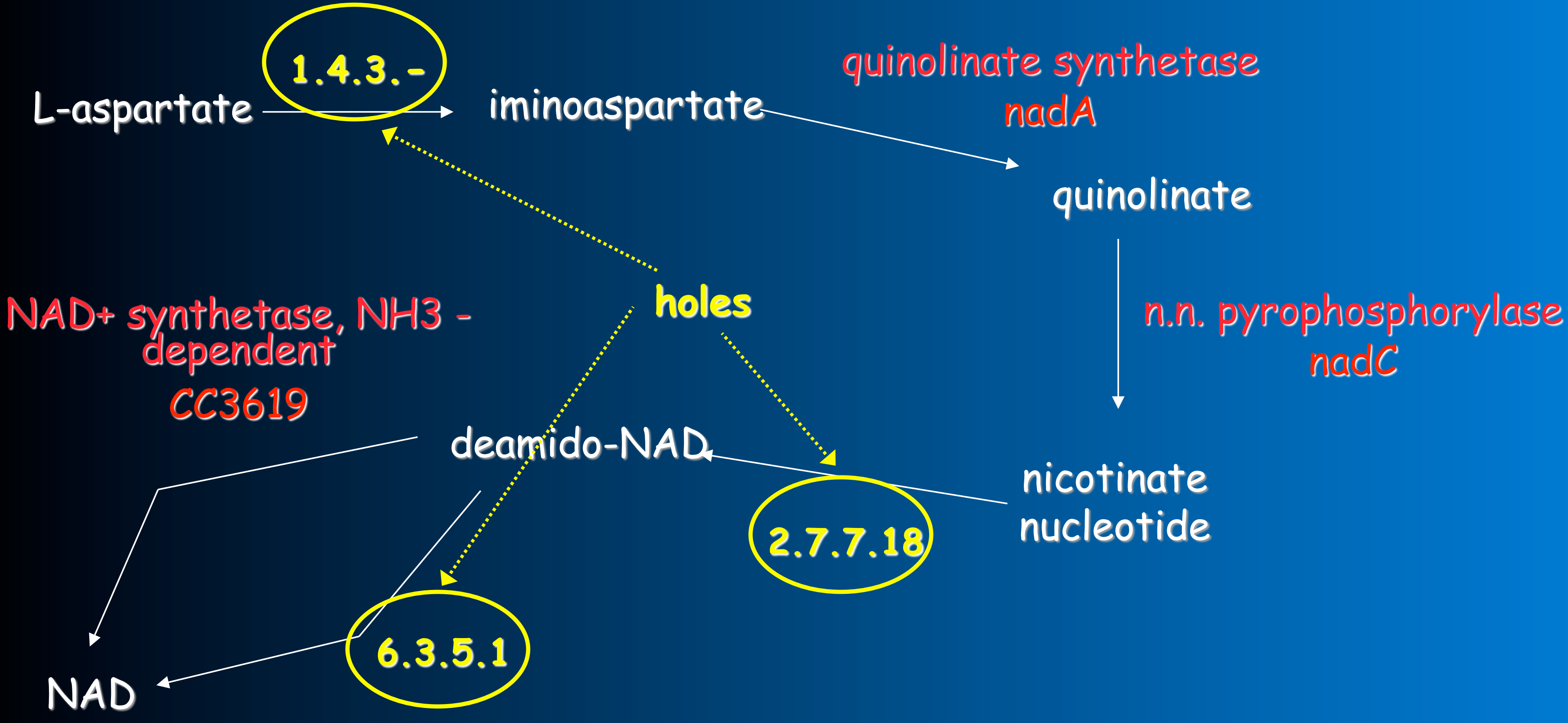
Pathway	Pathway Glyph	Total Rxns	Rxns Present in V. cholerae	Rxns Present in Other Pwys	Other Pwys
biosynthesis of proto- and siroheme		15	12	2	tRNA charging pathway cobalamin biosynthesis II (aerobic pathway)
biotin biosynthesis I		4	4	0	(none)
cobalamin biosynthesis I		7	6	2	cobalamin biosynthesis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

Pathway Hole Filling

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



L-aspartate

1.4.3.-

iminoaspartate

quinolinate synthetase
nadA

quinolinate

n.n. pyrophosphorylase
nadC

nicotinate nucleotide

2.7.7.18

deamido-NAD

holes

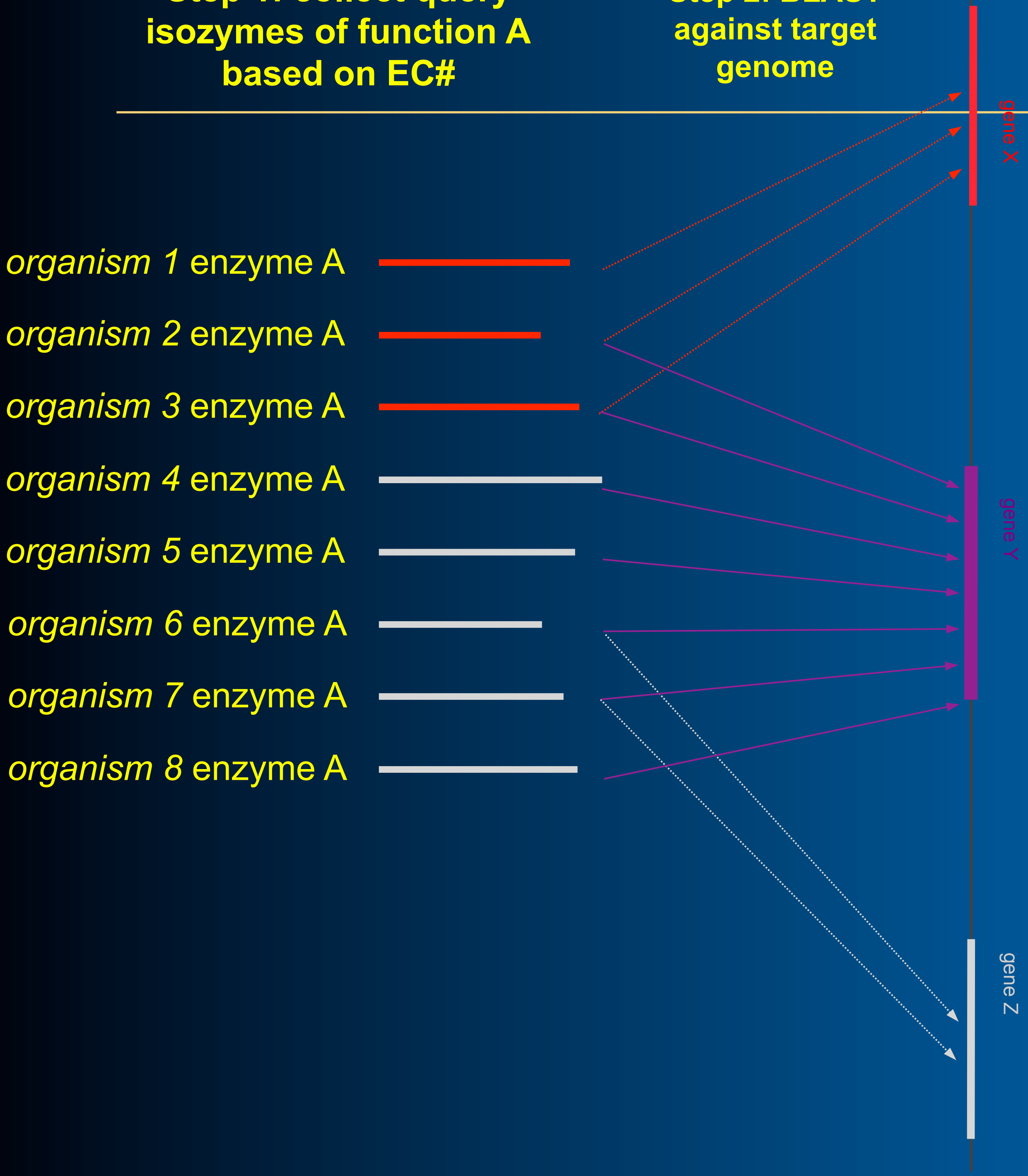
NAD+ synthetase, NH3 -dependent
CC3619

6.3.5.1

NAD

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

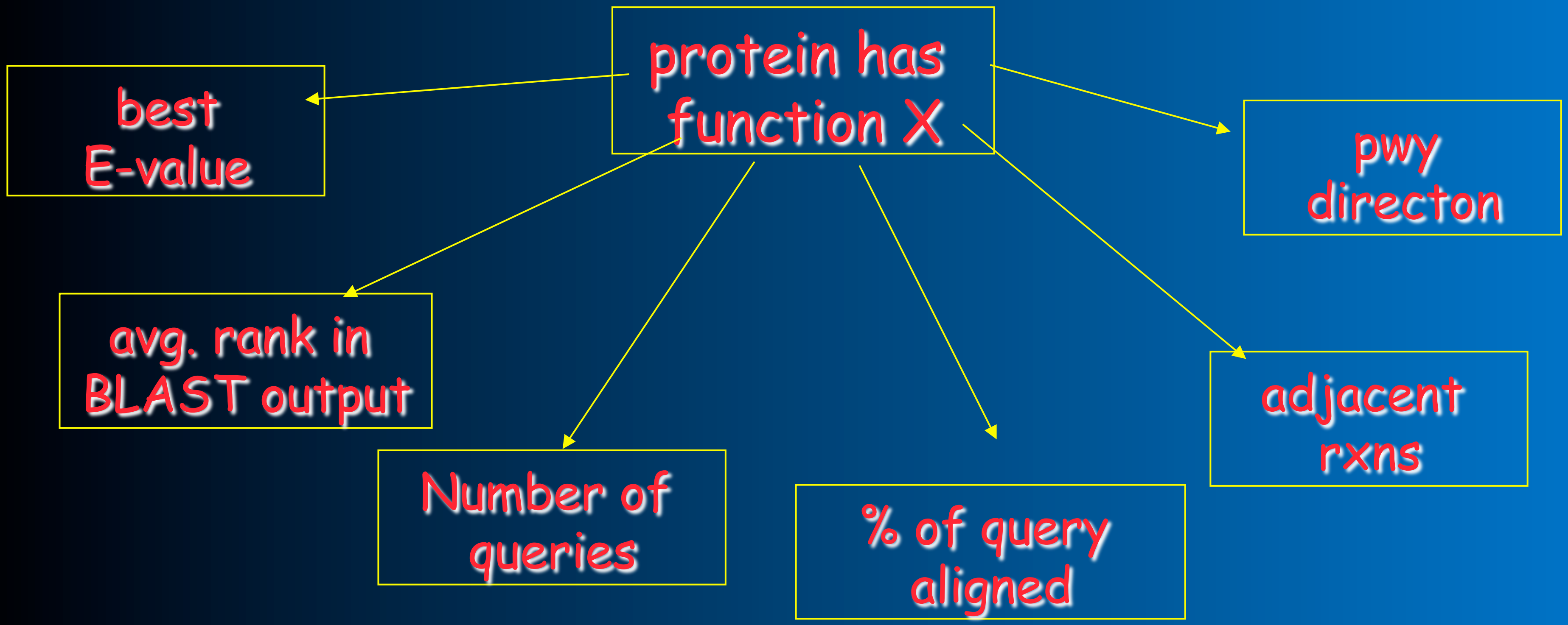
organism 8 enzyme A

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.)



protein has function X

best E-value

avg. rank in BLAST output

Number of queries

% of query aligned

adjacent rxns

pwy directon

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

- **In 90 minutes, I only got to here**
- **Included a 10-15 min demo**
- **3/8/2007 Brutlag class lecture**

PathoLogic Step 6: Build Cellular Overview Diagram

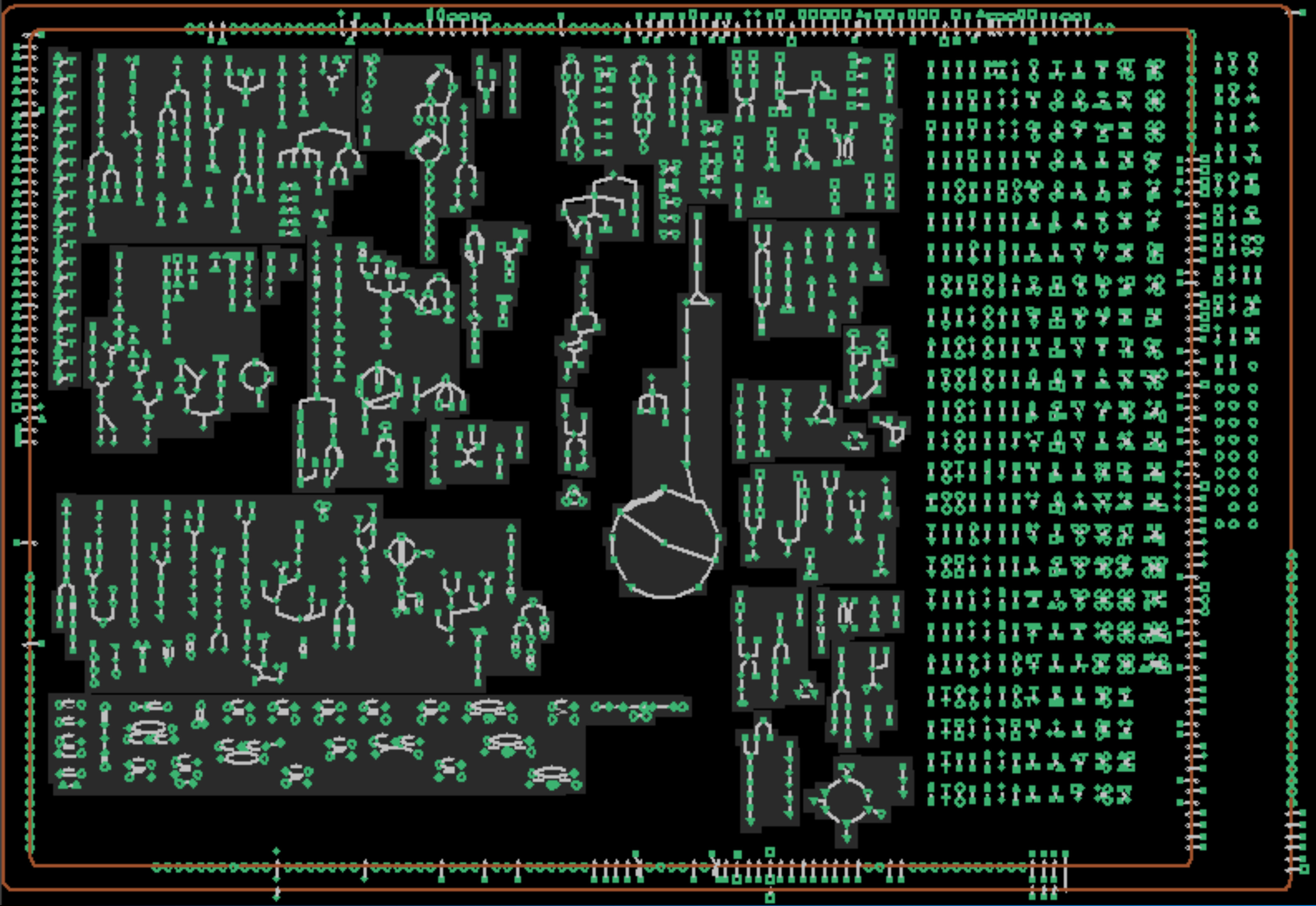
- Diagram encompassing metabolic, transport, and other cellular networks

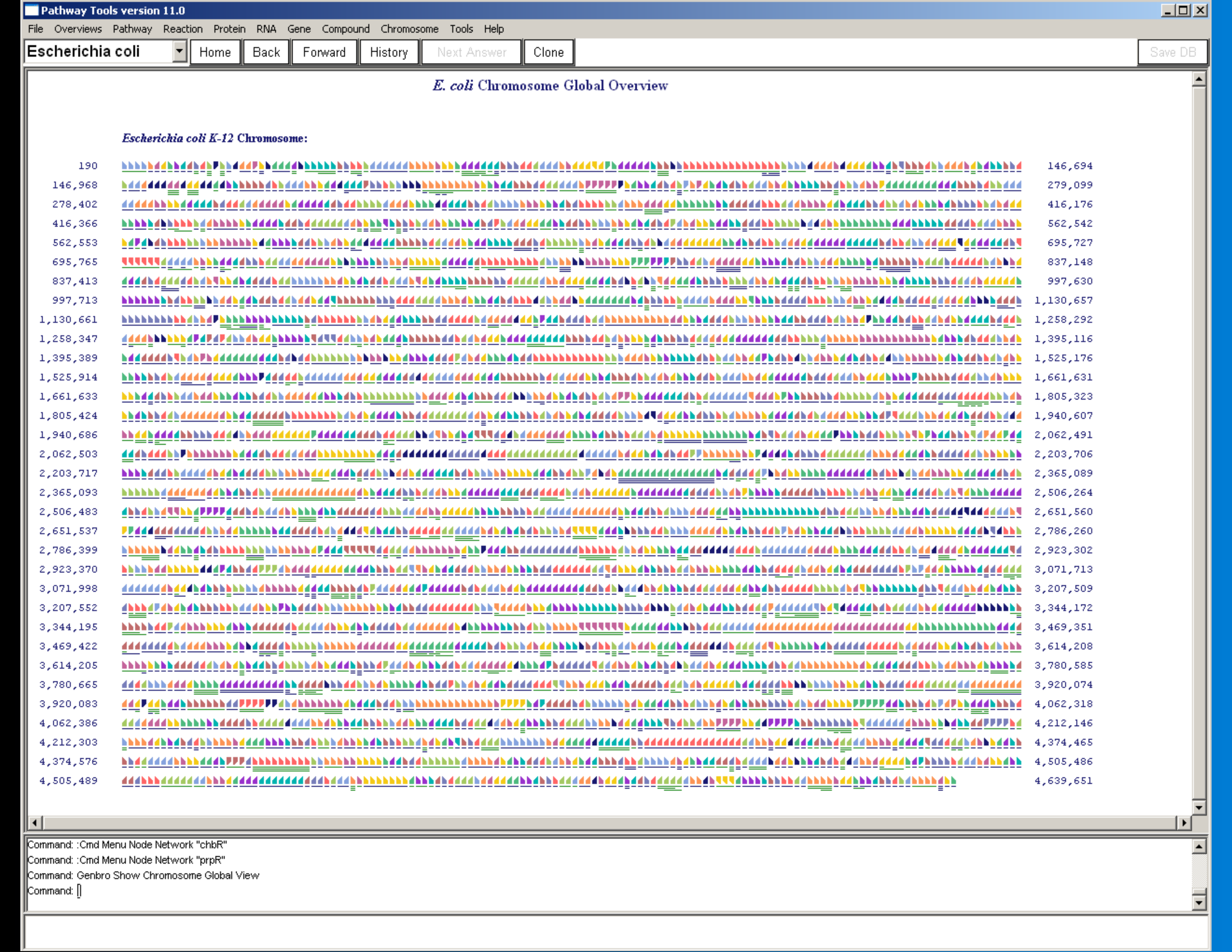
- Automatically generated for every BioCyc DB using advanced graph layout algorithm

- 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



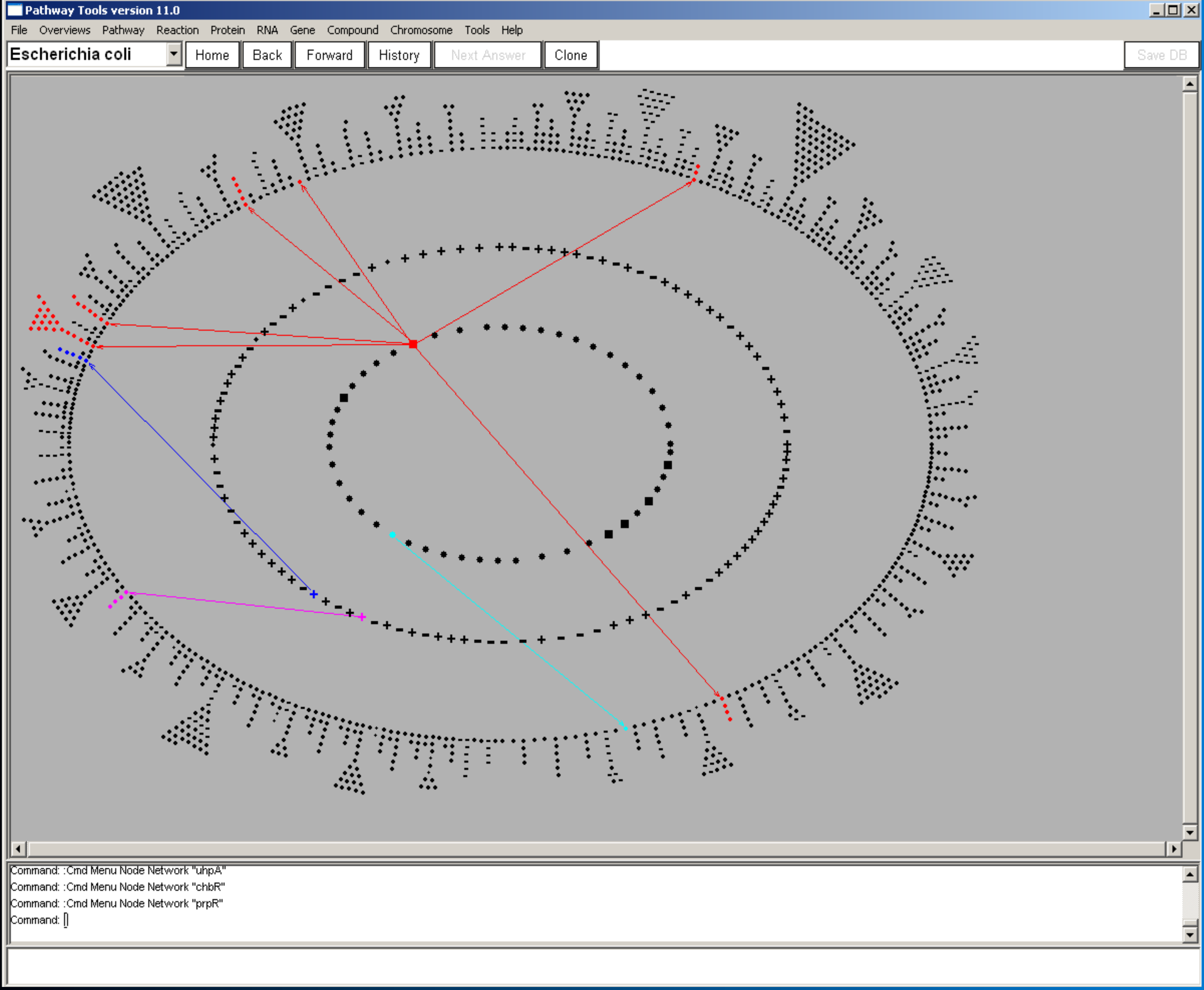


E. coli Chromosome Global Overview

Escherichia coli K-12 Chromosome:

190	146,694
146,968	279,099
278,402	416,176
416,366	562,542
562,553	695,727
695,765	837,148
837,413	997,630
997,713	1,130,657
1,130,661	1,258,292
1,258,347	1,395,116
1,395,389	1,525,176
1,525,914	1,661,631
1,661,633	1,805,323
1,805,424	1,940,607
1,940,686	2,062,491
2,062,503	2,203,706
2,203,717	2,365,089
2,365,093	2,506,264
2,506,483	2,651,560
2,651,537	2,786,260
2,786,399	2,923,302
2,923,370	3,071,713
3,071,998	3,207,509
3,207,552	3,344,172
3,344,195	3,469,351
3,469,422	3,614,208
3,614,205	3,780,585
3,780,665	3,920,074
3,920,083	4,062,318
4,062,386	4,212,146
4,212,303	4,374,465
4,374,576	4,505,486
4,505,489	4,639,651

```
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: []
```

Pathway Algorithms

- **Automated layout of metabolic pathways**
  - *Bioinformatics and Genome Research Conference* 1994 p225
- **Automated layout of cellular overview diagram**
- **Automated generation of metabolic map poster**

- **Forward propagation of metabolites through the metabolic network**
  - Consistency of a PGDB with respect to known growth-media requirements
  - *Pacific Symp Biocomputing* 2001:471
- **Identify dead-end metabolites**

- **Infer drug targets as choke points in metabolic network**
  - *Genome Research* 14:917 2004

Dead End Metabolites

- **Clues to extra/missing reactions**
- **A small molecule C is a dead-end if:**
  - (Def 1 easier to compute; Def 2 more accurate)
- **Definition 1:**
  - C is a substrate in only one reaction of the set of SMM reactions occurring in Compartment AND
  - No transporter acts on C in Compartment, nor on parent classes of C
- **Definition 2:**
  - 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

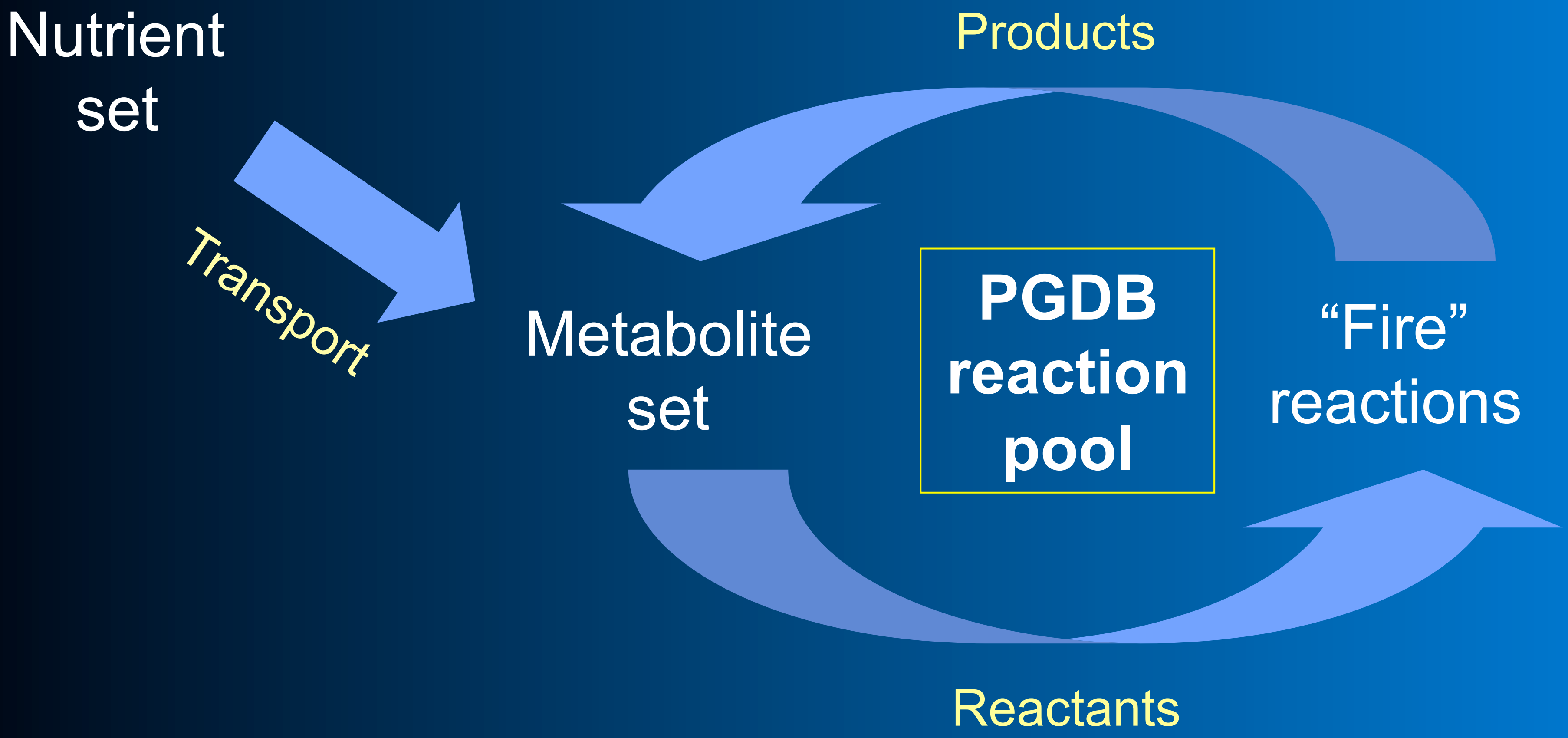
Global Consistency Checking of Biochemical Network

- **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
- **Can known growth medium yield known essential compounds?**

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

Algorithm: Forward Propagation Through Production System

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



Nutrient set

Transport

Metabolite set

Products

PGDB reaction pool

"Fire" reactions

Reactants

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)

Nutrients: A, B, C, E, F

$A + B \rightarrow W$

$C + D \rightarrow X$

$E + F \rightarrow Y$

$W + Y \rightarrow Z$

Produced Compounds: W, Y, Z

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 → 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

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**

Additional Pathway Tools Algorithms

- Predict metabolic pathway complement
- Automatic layout of Cellular Overview diagram
- Paint Omics datasets onto Cellular Overview
- Compare metabolic networks
- Reaction balance checker
- Chemical substructure search
- Predict operons
- Predict pathway hole fillers
- Qualitative path tracing from network inputs to network outputs