Acknowledgements (for yesterday's talk)

- **Bioinformatics Research Group**
- **DE-FG03-01ER63219 from the U.S. Department of Energy**

Conceptual differences between BioCyc and KEGG and their implications for (computational) biologists

Green, ML and Karp, PD. Genome Annotation Errors in Pathway Databases Due to Semantic Ambiguity in Partial EC Numbers. *Nucleic Acids Research 2005*, **33:**13, 4035-4039.

Green, ML and Karp, PD. The Outcomes of Pathway Database Computations Depend on Pathway Ontology. *Nucleic Acids Research,* accepted for publication.

Outline

- **KEGG**
- **BioCyc**
- **What differences exist?**
  - Treatment of partial EC numbers
  - Defining boundaries of pathways
  - (probably many others)
- **How do these differences impact users and the choice of database?**

KEGG

- 320 bacteria, 27 archaea, 17 eukaryotes
- 132 metabolic + 57 regulatory reference maps
- KO groups – each box in a KEGG pathway map represent a KO group
- Organism-specific maps generated by "painting" KO group annotations onto reference maps ; computed by orthology (ignores genome annotation)

BioCyc

- **MetaCyc – 759 metabolic pathways observed in specific organisms**
  - Identifies organism(s) where pathway has been confirmed.
  - Pathways from over 600 different organisms.
  - Includes data for reactions, compounds, genes, and proteins.
- **204 organism-specific PGDBs generated by matching** EC number **or reaction names to genome annotation.**

Assigning genes to reactions by EC number

- Most BioCyc reactions and KEGG KO groups have EC numbers
- Complete vs. partial EC numbers
  - Complete EC numbers (X.Y.Z.N) fully specify a reaction.
  - Partial EC numbers (X.Y.Z.-) do not fully specify a reaction.

Full EC number annotation

e.g., uxaC, EC# 5.3.1.12

uxaC catalyzes multiple reactions:

Galacturonate <=> D-tagaturonate

Glucoronate <=> Fructoronate

Interpreting Partial EC numbers

- BioCyc – gene-reaction assignments based on full EC numbers or name matching
- KEGG – genes assigned to KO groups by orthology (partial and full EC numbers)

Problem:

What reaction is indicated by a partial EC number?

Partial EC number annotation

e.g., caiB, EC# 2.8.3.-

caiB catalyzes:

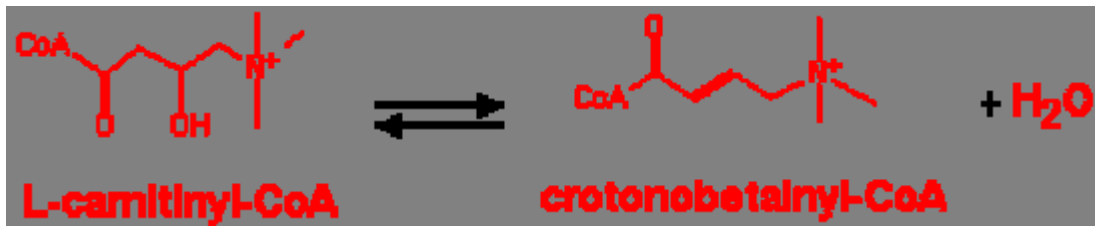
L-carnitine + γ-butyrobetainyl-CoA

<=> γ-butyrobetaine + L-carnitinyl-CoA

But the reaction cannot be determined by caiB's EC#!

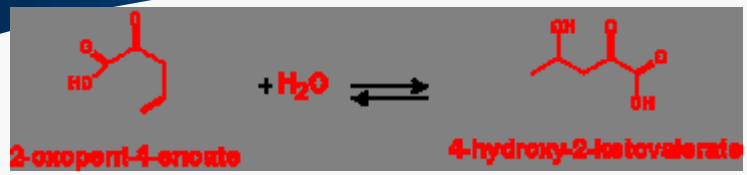
Two meanings of partial EC numbers

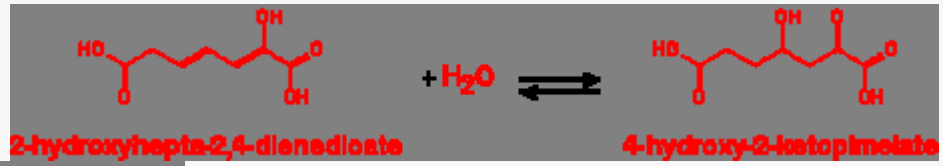
EC# 4.2.1.-: “This enzyme is a novel hydro-lyase, catalyzing a specific reaction, but the NC-IUBMB has not yet assigned an EC number” (full EC number not available yet).

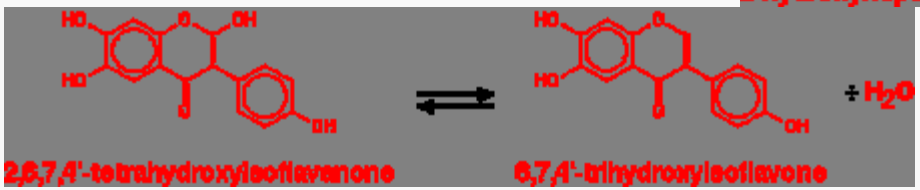


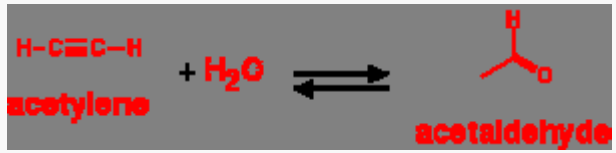
Two meanings of partial EC numbers

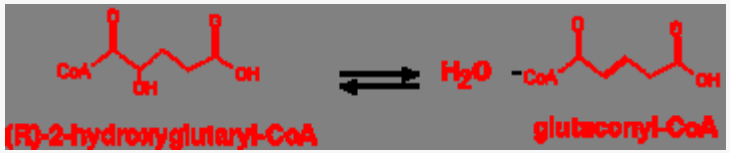
EC# 4.2.1.-: “This enzyme is a hydro-lyase, but I do not know its substrate specificity” (full EC number unknown)











How do these differences impact pathway databases?

- Surveyed examples from KEGG and BioCyc
- Retrieved each set of genes all assigned partial EC number X
- Looked for systematic errors due to assigning genes to multiple reactions with the same partial EC number
- Compared to UniProt/EcoCyc or likelihood of correctness based on the list of reactions

Example: EC# 4.2.1.-

- **KEGG**
  - appears in 10 pathway maps
  - b0036 and b1517 assigned to 16 distinct reactions
- **EcoCyc:**
  - b0036: carnitine racemase (EC 4.2.1.-); crotonobetainyl-CoA hydratase / carnitine racemase
  - b1517: putative aldolase (EC 4.2.1.-)

ENTRY b0036 CDS E.coli
NAME caiD
DEFINITION carnitinyl-CoA dehydratase [EC:4.2.1.-]

ENTRY b1517 CDS E.coli
NAME yneB
DEFINITION hypothetical 31.9 kD protein in hipB-uxaB intergenic region [EC:4.2.1.-]

Perillyl-CoA + H2O <=> 2-Hydroxy-4-isopropenylcyclohexane-1-carboxyl-CoA
(3R)-3-Isopropenyl-6-oxoheptanoate + CoA + ATP <=> (3R)-3-Isopropenyl-6-oxoheptanoyl-CoA + H2O + AMP + Pyrophosphate
Glutaconyl-1-CoA + H2O <=> 2-Hydroxyglutaryl-CoA
Cyclohex-1-ene-1-carboxyl-CoA + H2O <=> 2-Hydroxycyclohexane-1-carboxyl-CoA
E-Phenylitaconyl-CoA + H2O <=> (Hydroxymethylphenyl)succinyl-CoA
6-Hydroxycyclohex-1-enecarbonyl-CoA + H2O <=> 2,6-Dihydroxycyclohexane-1-carboxyl-CoA
6-Carboxyhex-2-enoyl-CoA + H2O <=> 3-Hydroxypimeloyl-CoA
Acetaldehyde <=> Acetylene + H2O
Homocystine + 2 Cyanide <=> alpha-Amino-gamma-cyanobutanoate + Homocysteine + Thiocyanate
4-Hydroxy-4-methyl-2-oxoglutarate <=> 4-Carboxy-2-oxo-4-pentanoate + H2O
4-Carboxy-4-hydroxy-2-oxoadipate <=> 4-Carboxy-2-hydroxy-cis,cis-muconate + H2O
2-Oxohept-3-enedioate + H2O <=> 4-Hydroxy-2-oxo-heptandioate
2-Oxohept-3-enedioate + H2O <=> 2,4-Dihydroxyhept-2-enedioate
4-Hydroxyphenylacetonitrile <=> 4-Hydroxyphenylacetaldoxime
3alpha,7alpha,12alpha,24-Tetrahydroxy-5beta-cholestanoyl-CoA <=> 3alpha,7alpha,12alpha-Trihydroxy-5beta-cholest-24-enoyl-CoA + H2O
3alpha,7alpha-Dihydroxy-5beta-cholest-24-enoyl-CoA + H2O <=> 3alpha,7alpha,24-Trihydroxy-5beta-cholestanoyl-CoA

Systematic analysis

Group of Genes	Number of genes	
All *E. coli* genes in KEGG	**4411**	
With EC#s (full or partial)	**869**	
With partial EC#'s	**135**	**% of parial Ecs**
correct	**38**	**28.1**
incorrect	**59***	**43.7**
unfinished/missing	**14**	**10.4**
no associated reaction	**21**	**15.6**
undetermined	**3**	**2.2**

* About half of the instances we identified in KEGG have been updated (35 genes no longer in pathways)

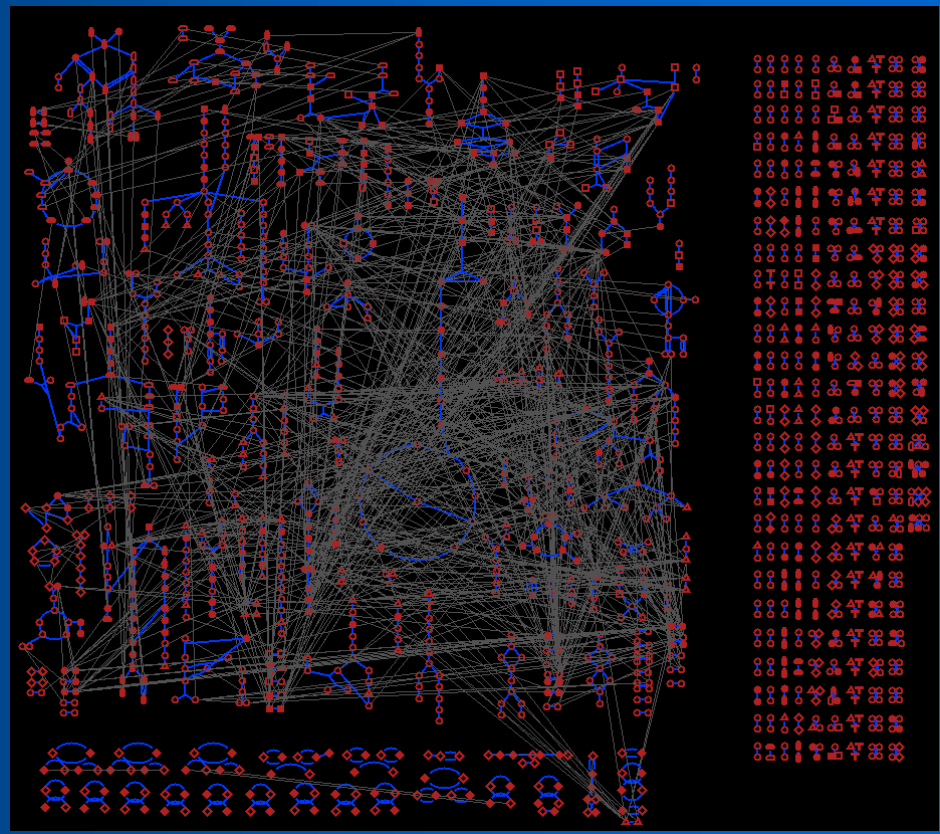
Recommendation for specification of partial EC numbers

- When the reaction is unknown use EC number of the form "**X.Y.Z.?**"
- When the reaction is known, but EC not yet supplied by NC-IUBMB use EC number of the form "**X.Y.Z.n**"

- Genome annotation experts can be explicit about which meaning they intend when assigning a partial EC number to a gene.

The Metabolic Network

- A large network of interconnected biochemical reactions
- In this state, not very useful for a person
- Pathways segment the network into usable components
- ***How are pathway boundaries defined?***



KEGG rules for pathway boundaries

- Each map can encompasses mutually exclusive processes (i.e., biosynthesis and degradation of a metabolite)
  - Thus, not regulated as a unit (mutually exclusive parts)
  - Tend to be substrate-centric
- Reference maps combine pathways from multiple organisms
  - No consideration of evolutionary conservation of modules
  - Integrate “all possible” transformations

BioCyc criteria for pathway boundaries

- Goal: define pathways that correspond to…
  - distinct biological processes
  - conserved, functional, atomic modules of the network
- (Older pathways in BioCyc may not consistently reflect these rules)

Rules for BioCyc pathway boundaries

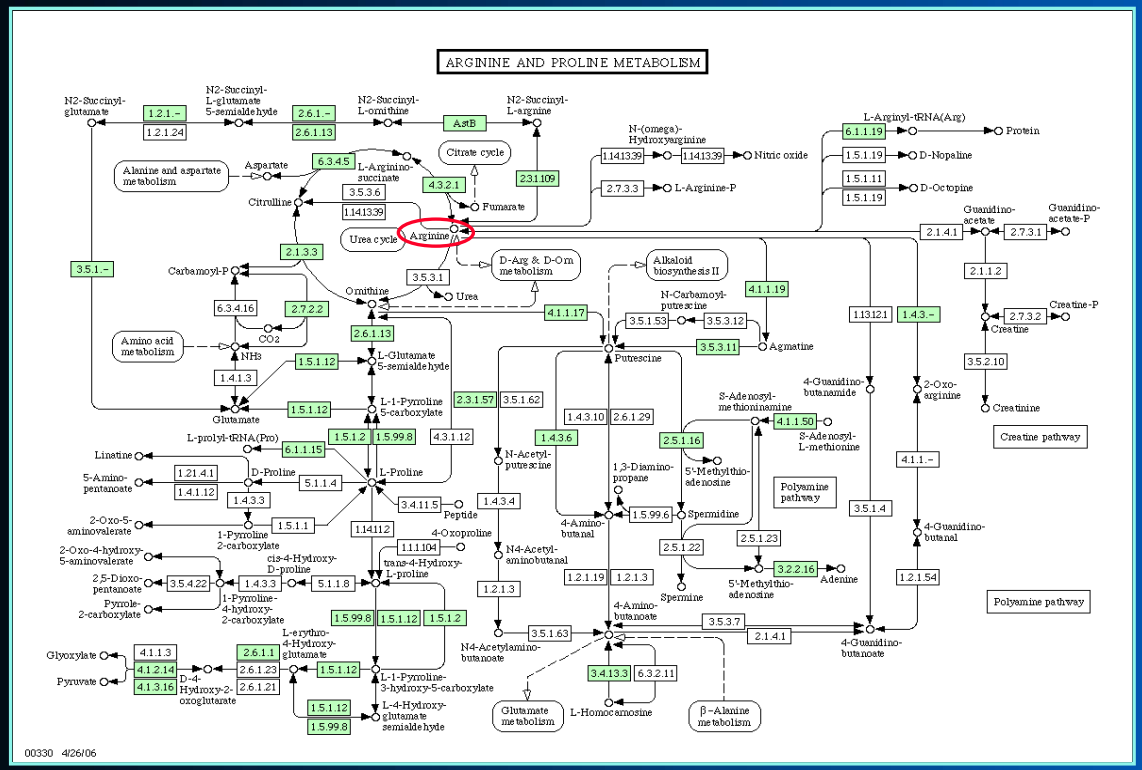
1. Find a single common biological process
2. Define boundaries at high-connectivity substrates
3. Pathway reactions share common regulation
4. Define boundaries at stable substrates
5. Pathways exhibit evolutionary conservation

Rules for BioCyc pathway boundaries

Find a common biological process

- e.g., degradation of arginine
- most reactions are active simultaneously

KEGG Arginine and Proline Metabolism



Corresponding BioCyc pathways:

1. Arginine biosynthesis I
2. Arginine degradation II (glutamate & succinate)
3. Arginine degradation III (putrescine)

Rules for BioCyc pathway boundaries

Define boundaries at high-connectivity substrates (HCS)

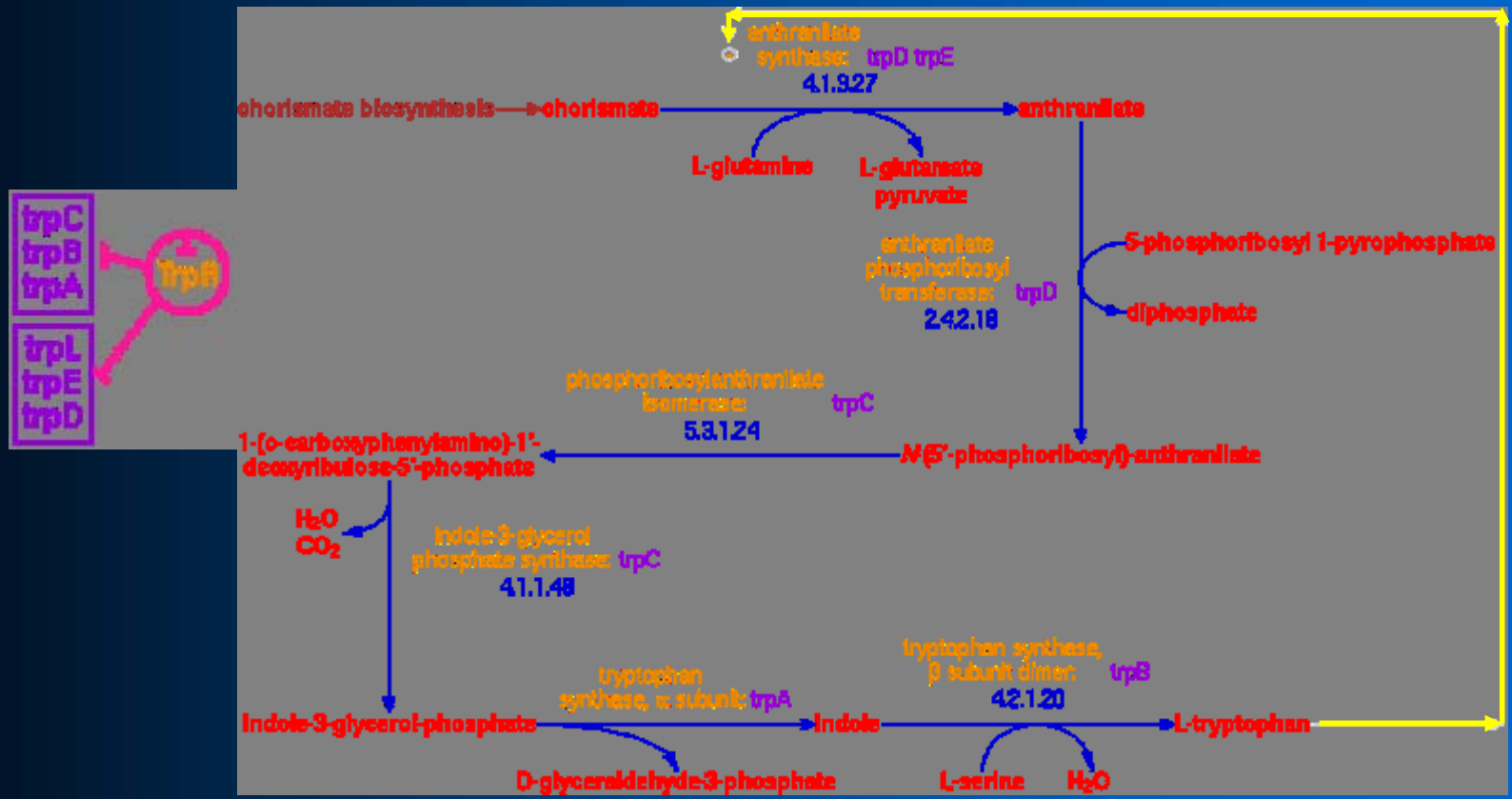
- Most common HCSs are metabolites of central metabolism (glycolysis, TCA cycle, pentose phosphate pathway)
- BioCyc biosynthesis pathways start at HCS and end at building block macromolecule/cofactor/coenzyme.
- BioCyc catabolic pathways typically end at HCS

glucose-6-phosphate, fructose-6-phosphate, ribose-5-phosphate, erythrose-4-phosphate, triose phosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate, acetyl CoA, alpha-oxoglutarate, succinyl CoA, oxaloacetate, and sedoheptulose-7-phosphate

Rules for BioCyc pathway boundaries

Pathway reactions share common regulation

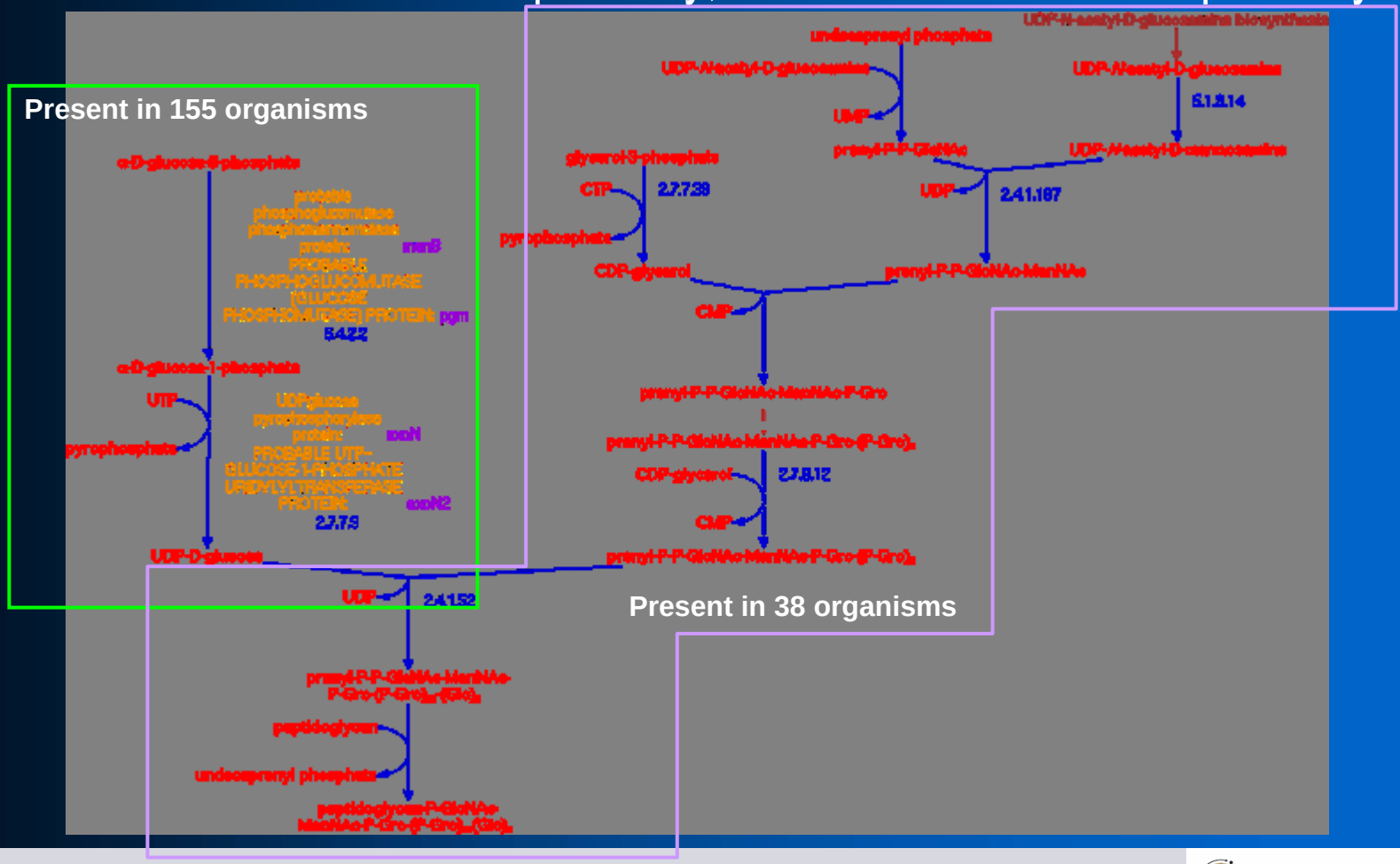
- Substrate-level enzyme inhibition/activation
- Expression-level repression/attenuation



Rules for BioCyc pathway boundaries

Pathway exhibit evolutionary conservation

- Given the difference in the level of conservation of the two branches of this pathway, it was divided into two pathways



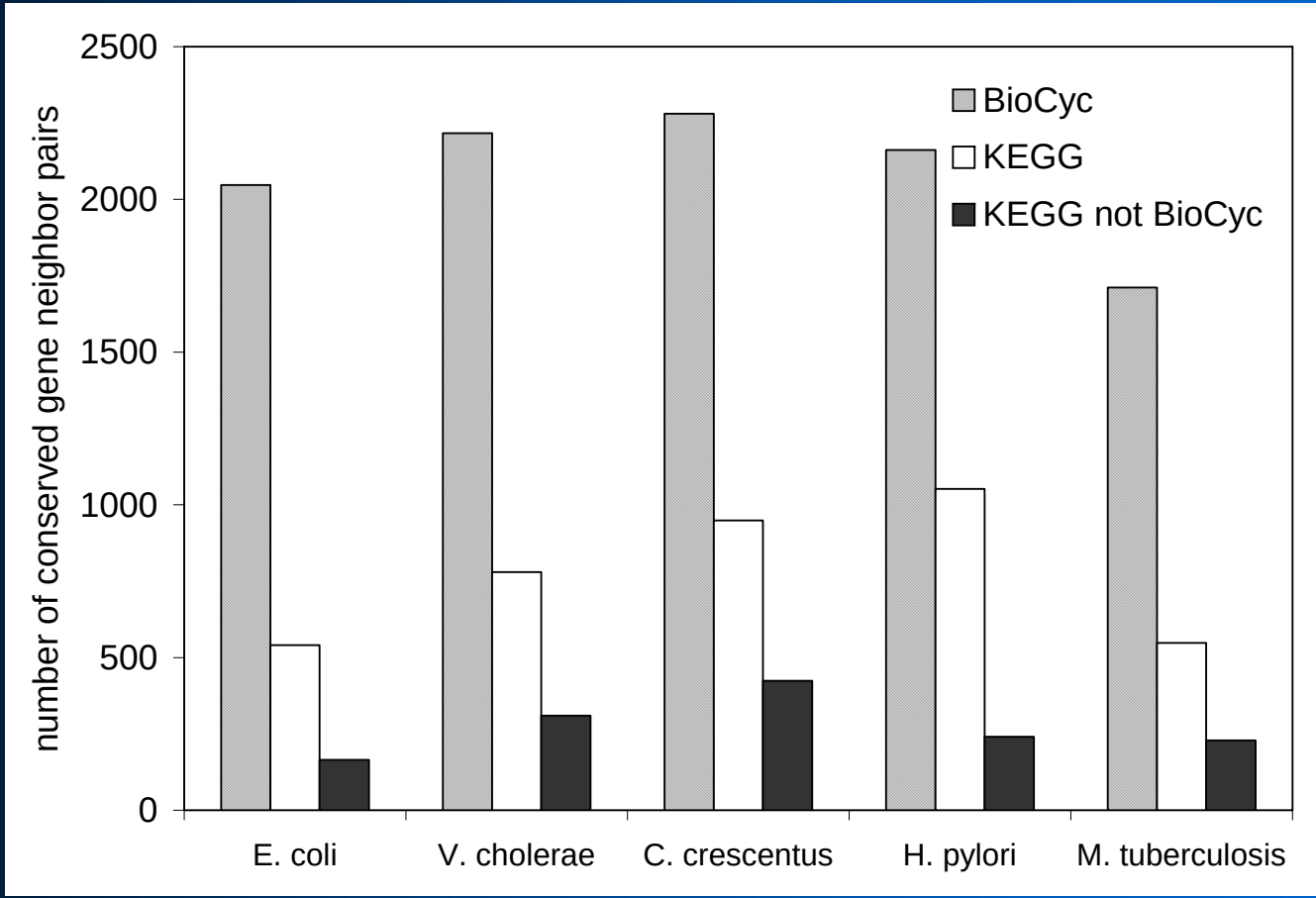
Global Properties of Pathways in BioCyc and KEGG

- **Surveyed functional relatedness of gene pairs from same pathway**
- 10000 random gene pairs from same EcoCyc pathway or same KEGG map
- Count # pairs related by:
  - Conserved neighbors
  - Similar phylogenetic profiles
  - Gene cluster (operon)
  - Gene fusion

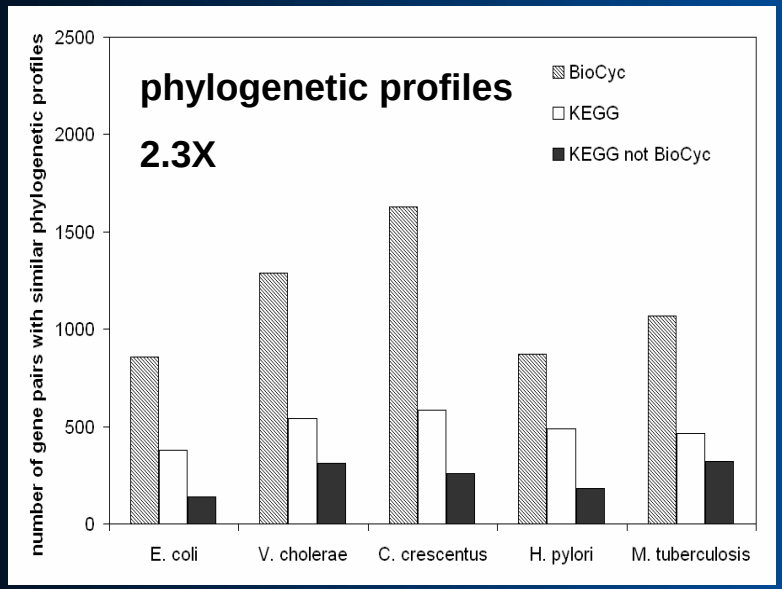
  (all genome context data from Prolinks)

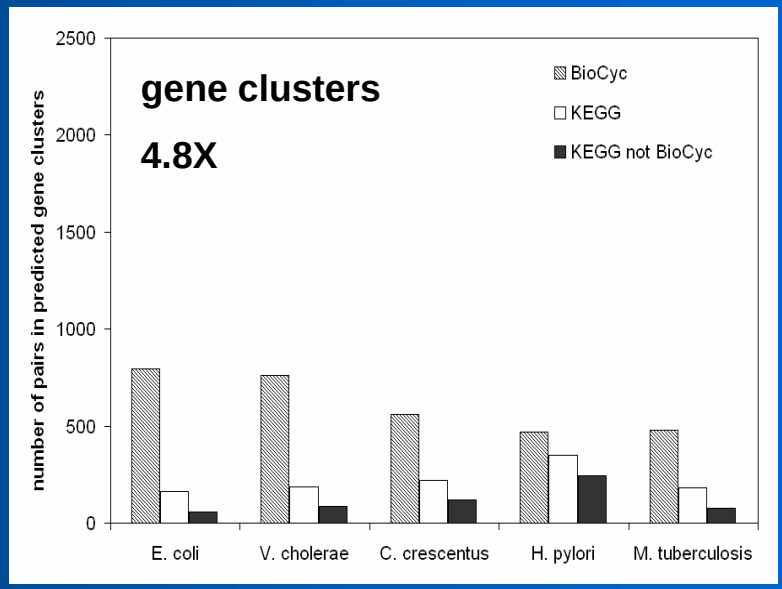
Conserved gene neighbors…

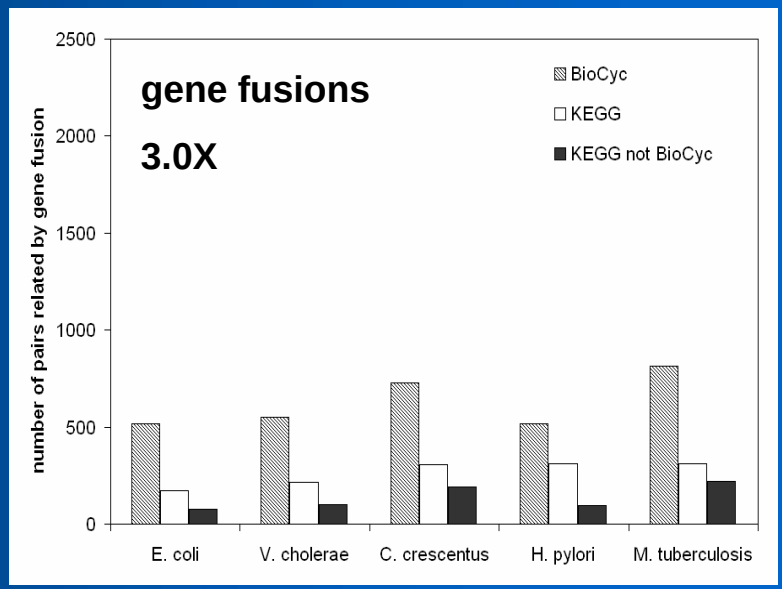
A pair of genes from the same EcoCyc pathway is **3.8 times** more likely to be related than from the same KEGG *E. coli* pathway.



Results for other genome context methods are similar







Summary of genome context methods

- Genome context data show that BioCyc pathways better represent functionally related groups of genes.

So, which database should you use?

Each DB is better suited for different tasks

- Encyclopedia of distinct metabolic processes present in a given organism
  - BioCyc pathways represent reactions ocurring in one organism in one biological process
  - KEGG integrates multiple processes in one pathway; unclear indication of what is present in an organism
- Encyclopedia of processes impinging on a given substrate.
  - Each KEGG map combines many processes related to a given substrate within one diagram; easily accessible
  - BioCyc compound pages, superpathways, cellular overview can provide summary

Each DB is better suited for different tasks

- Pathway prediction or reconstruction, and detection of missing pathway components
  - BioCyc – easier identification of false positive predictions based on evidence
  - KEGG – easier to investigate pathway variants on one pathway map
- Gold standard for developing methods that predict pathways and for genome context methods
  - BioCyc – close range relationships; evolutionarily conserved modules
  - KEGG – perhaps more general relationships related to action on a common substrate
  - Carefully consider the type of relationships learned from data compared to desired predictions

Each one better suited for different tasks

- Gold standard for developing genome context methods
  - Computational researchers should carefully consider the type of relationships being learned from data compared to desired predictions
- Analysis of “-omics” datasets
  - No clear advantage
  - Probably dependent on range of relationships desired

Acknowledgements

- Bioinformatics Research Group
- RR07861 and GM70065 from the National Institutes of Health
- DE-FG03-01ER63219 from the U.S. Department of Energy

The end.

Considerations for using KEGG

- Advantages
  - Provide a big picture of reactions related to a specific substrate (although PGDB overview can do this too)
- Disadvantages
  - Inconsistent functions for genes
    - annotation ≠ assigned reactions
  - Incorrect function assignments
    - training and validating methods using bad data
    - a single incorrectly assigned gene (by partial EC) results in multiple incorrect associations (N = # genes assigned to pathway)
  - Overestimates the number of related gene pairs.