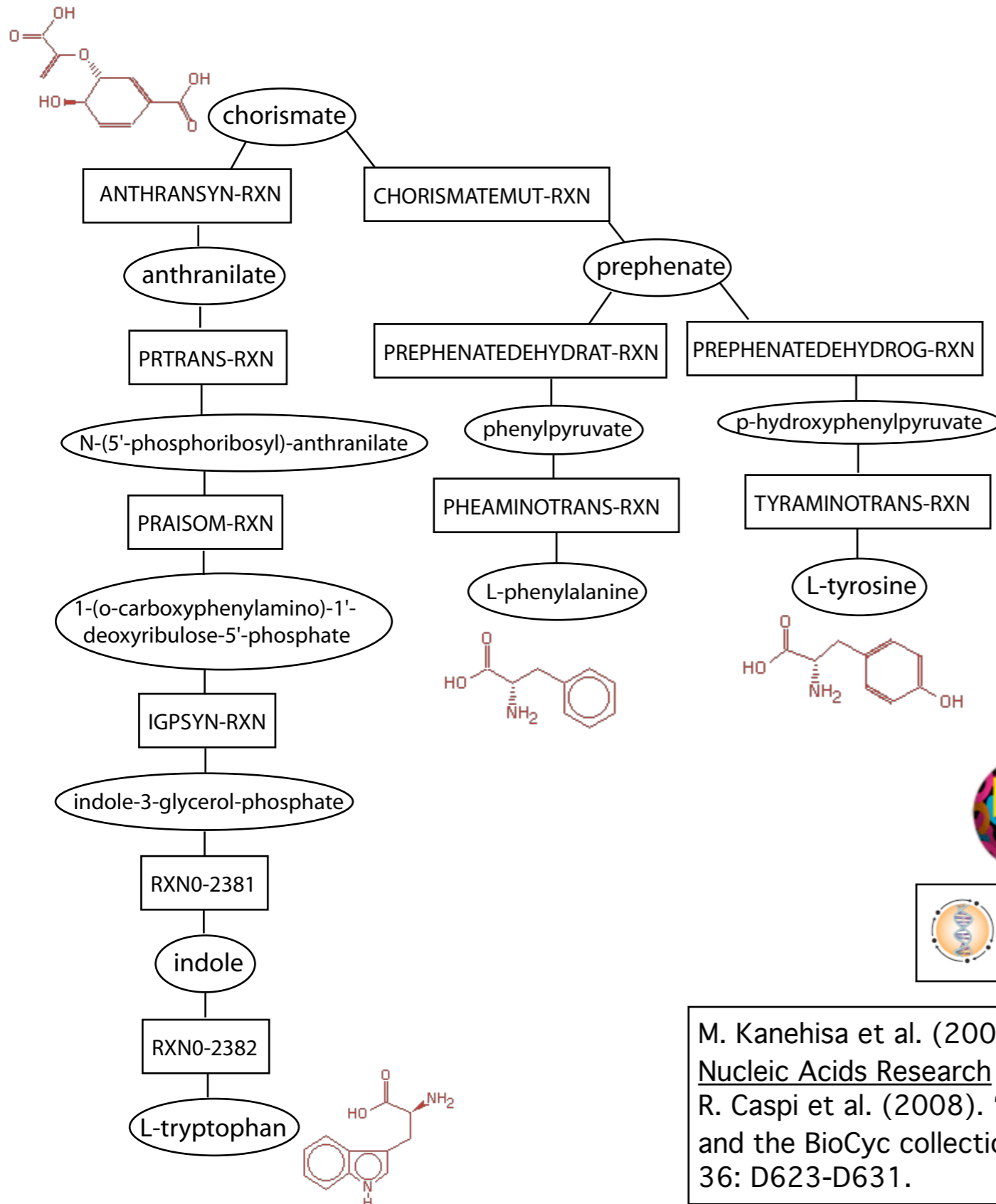


Metabolic Pathway Inference using Random Walks and Shortest-Paths Algorithms

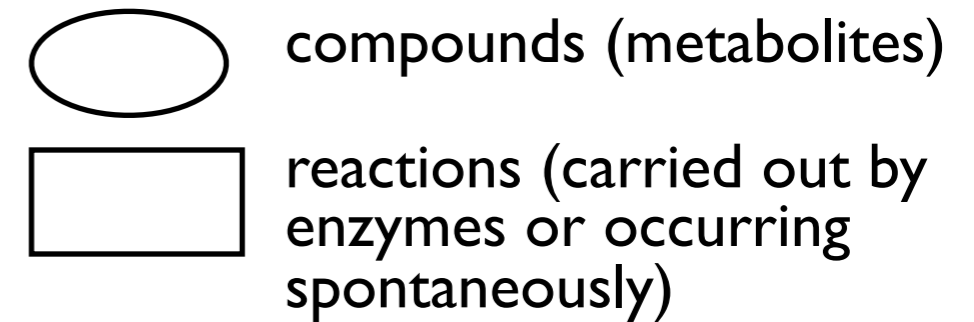
Karoline Faust, Jérôme Callut, Pierre Dupont, Jacques van Helden

Biological background

part of aromatic amino acid biosynthesis in *E. coli* (BioCyc)



metabolic data:



pathways (annotated combinations of reactions and compounds)

sources for metabolic data:

biochemical textbooks

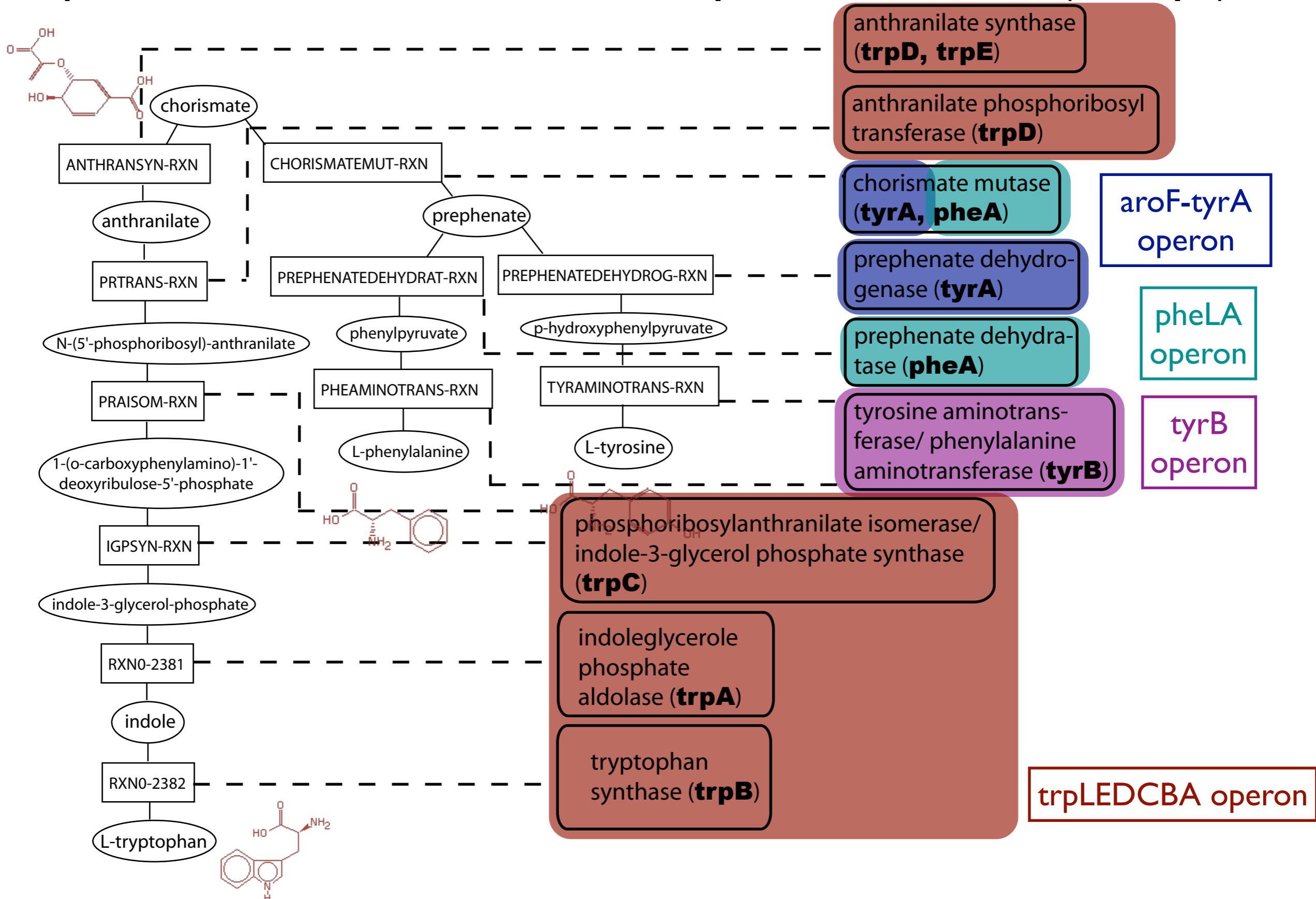
metabolic databases



M. Kanehisa et al. (2008). "KEGG for linking genomes to life and the environment.", *Nucleic Acids Research* 36: D480-D484.
R. Caspi et al. (2008). "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." *Nucleic Acids Research* 36: D623-D631.

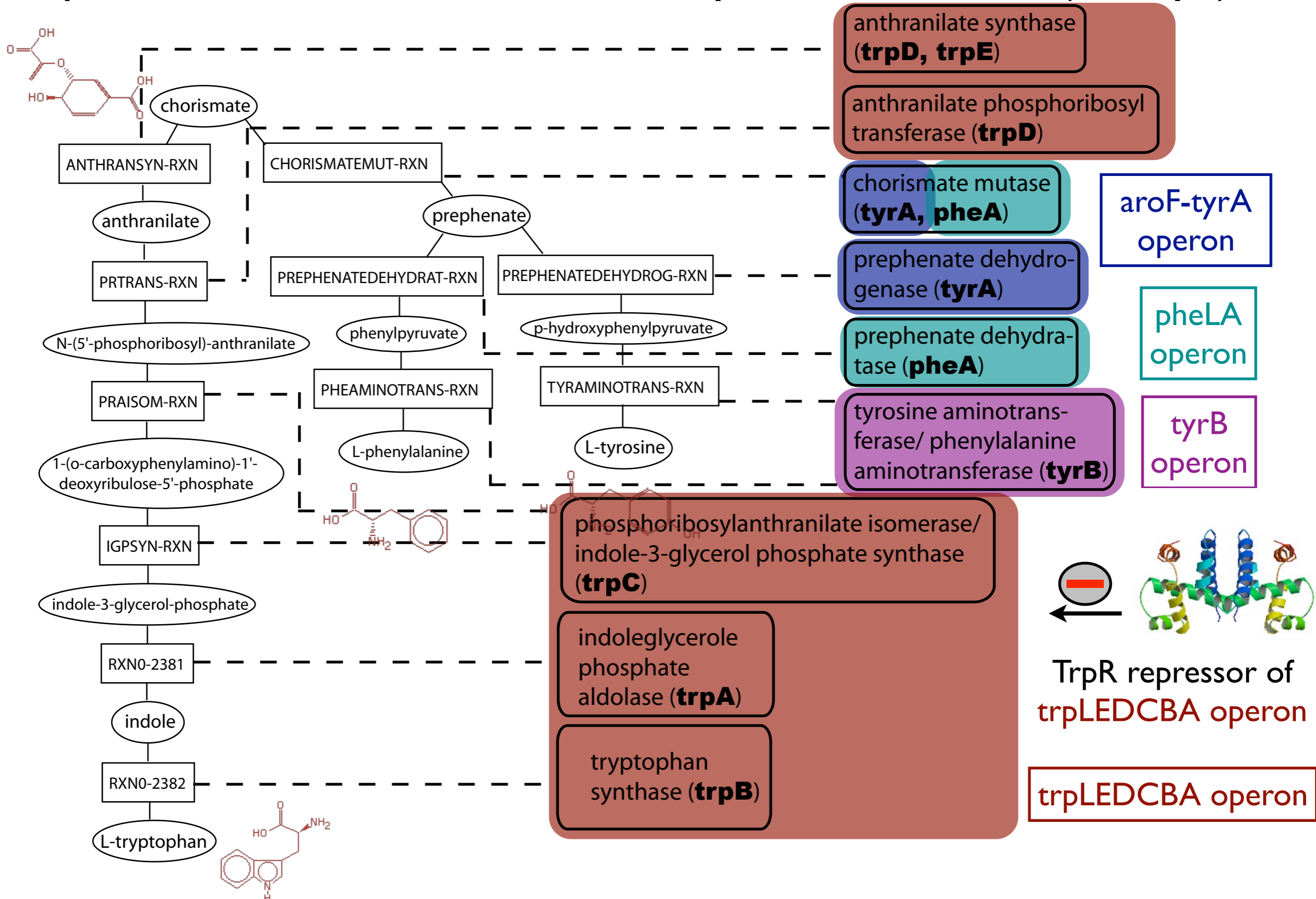
Biological background

part of aromatic amino acid biosynthesis in *E. coli* (BioCyc)



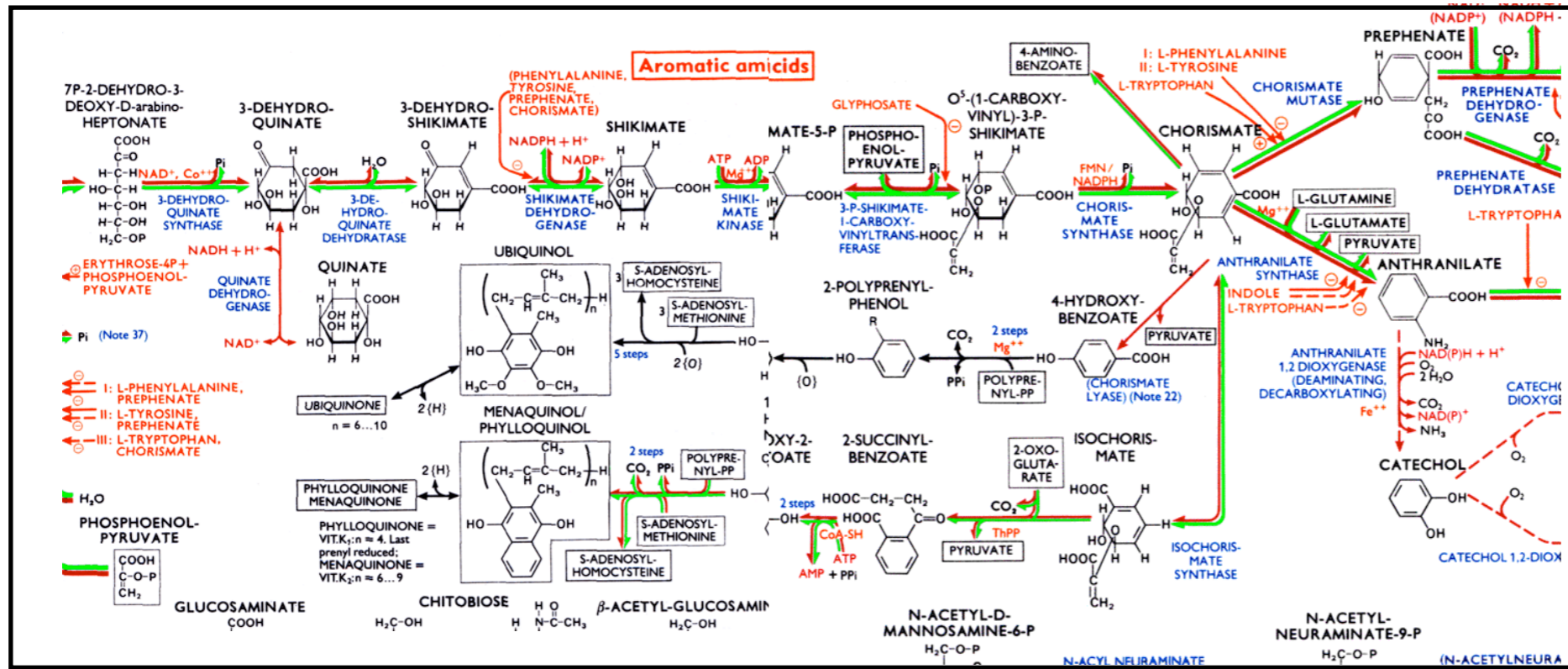
Biological background

part of aromatic amino acid biosynthesis in *E. coli* (BioCyc)

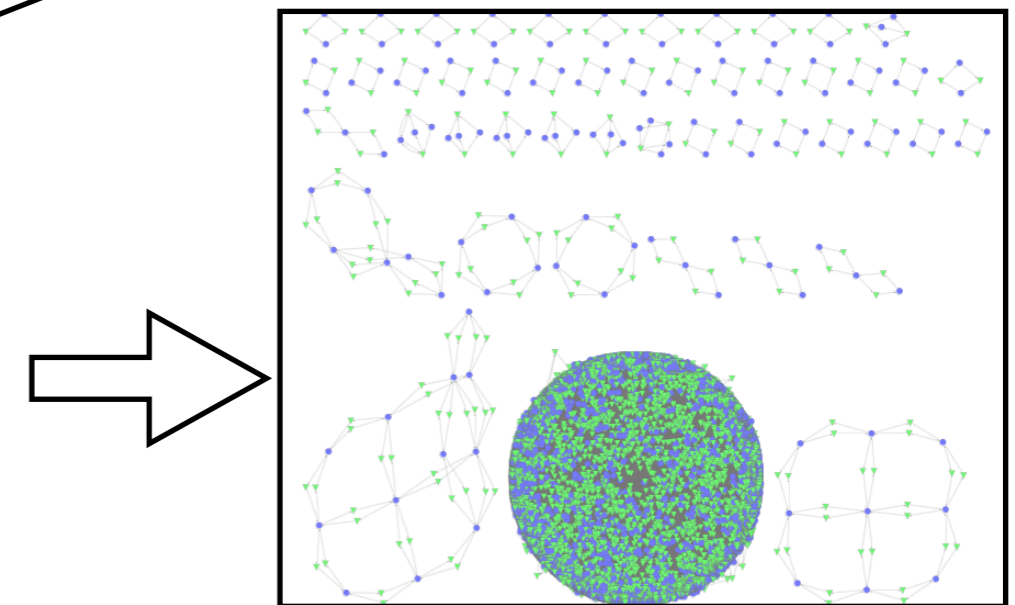
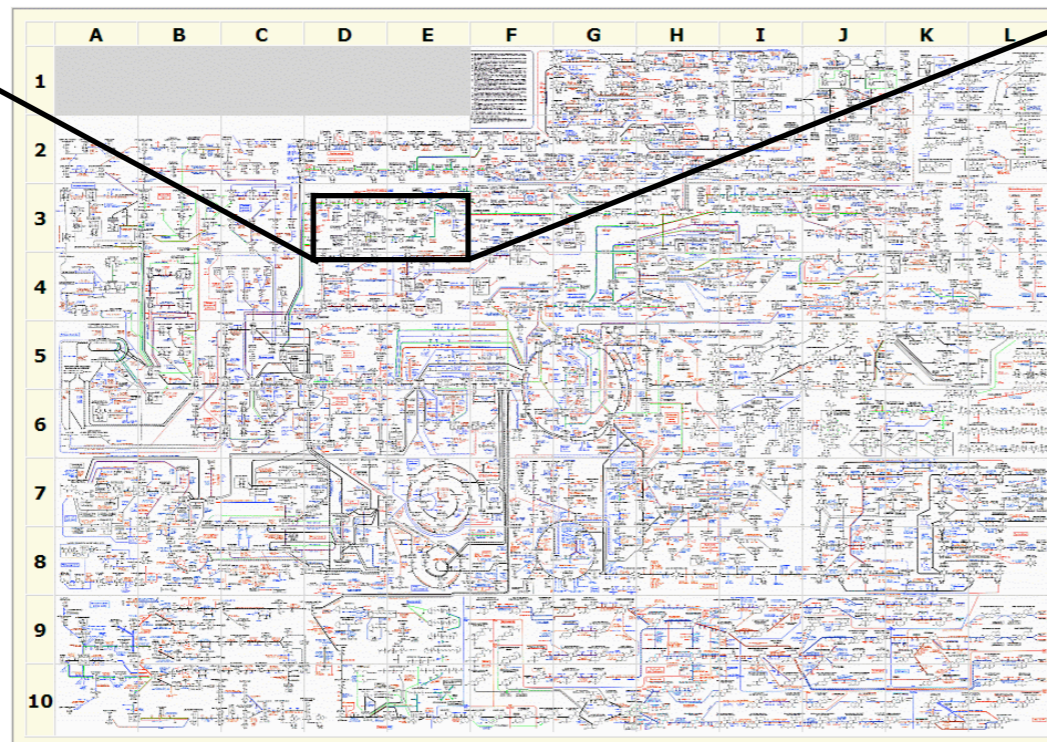


Biological background

Building a metabolic network from all known reactions



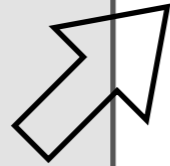
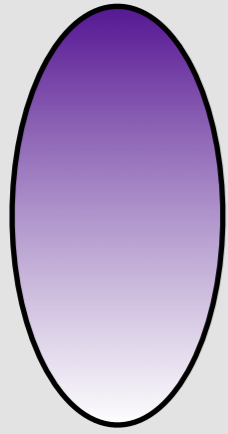
metabolic data can be represented in form of bipartite graphs consisting of compound and reaction nodes



metabolic graph

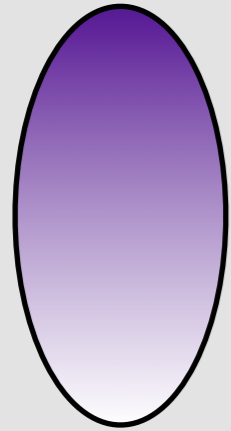
biochemical pathway wall chart (Roche)

Biological question



organism with
annotated genes
and unknown
metabolism

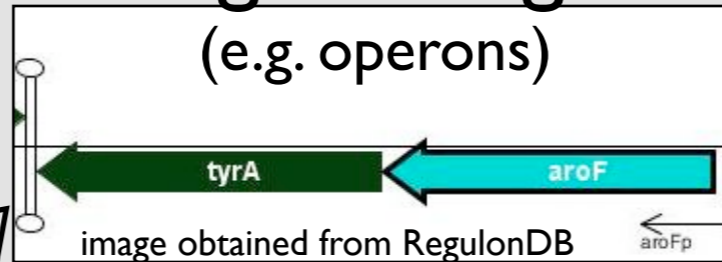
Biological question



organism with annotated genes and unknown metabolism

co-regulated genes

(e.g. operons)



co-expressed genes
(obtained from micro-array data)

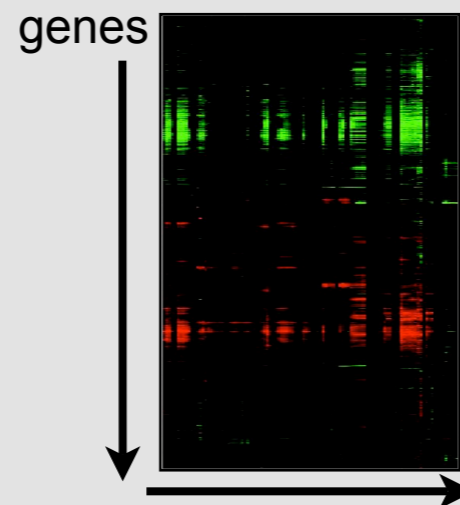
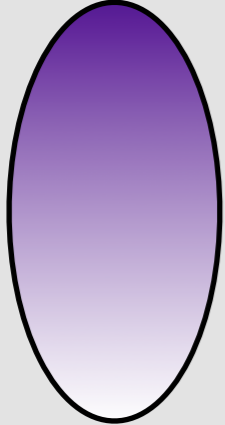


image taken from Gasch et al.,
Mol. Biol. of the Cell, 2000

Biological question



organism with annotated genes and unknown metabolism

co-regulated genes
(e.g. operons)

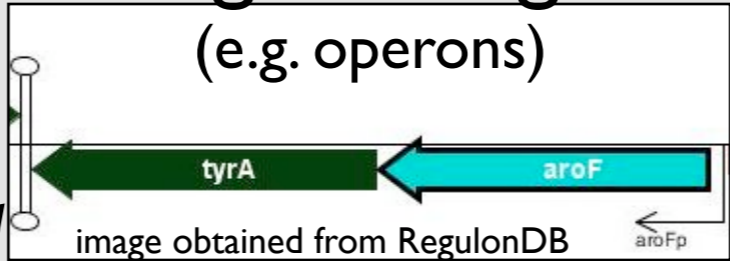
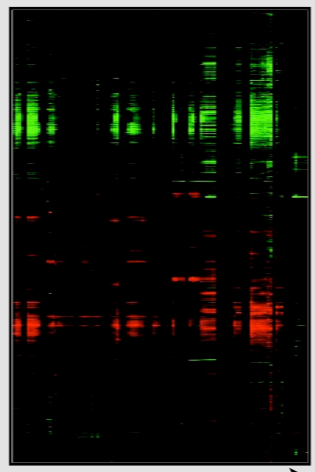


image obtained from RegulonDB

co-expressed genes
(obtained from micro-array data)



genes

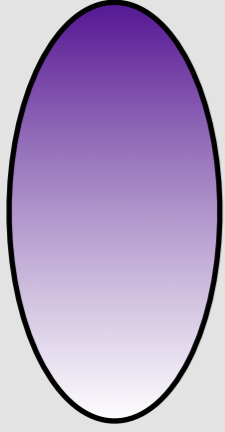
conditions

image taken from Gasch et al., Mol. Biol. of the Cell, 2000

gene group of interest

- Gene A
- Gene B
- Gene C
- Gene D

Biological question



organism with annotated genes and unknown metabolism

co-regulated genes (e.g. operons)

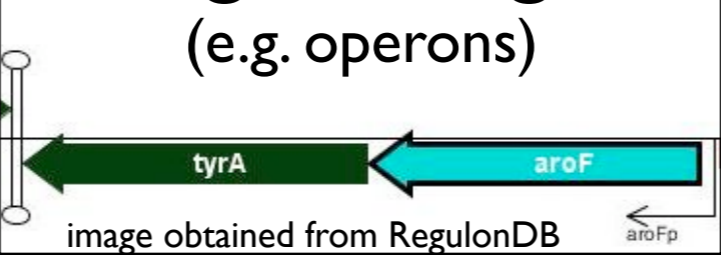
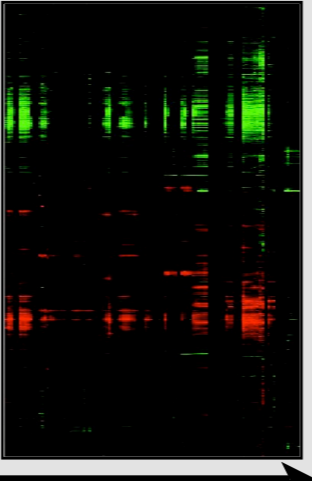


image obtained from RegulonDB

co-expressed genes (obtained from micro-array data)



genes

conditions

image taken from Gasch et al., Mol. Biol. of the Cell, 2000

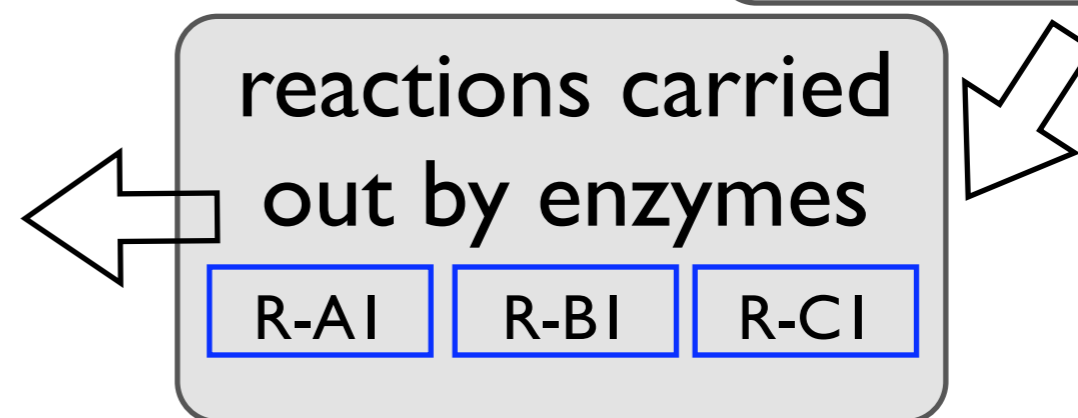
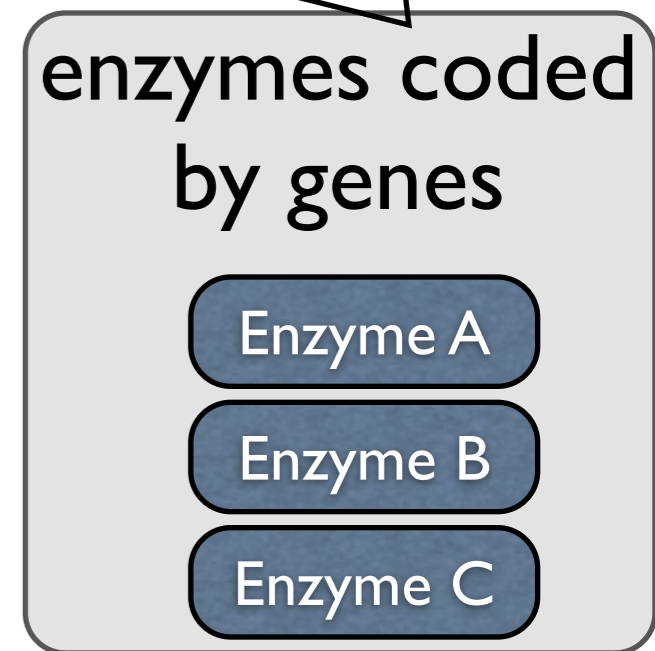
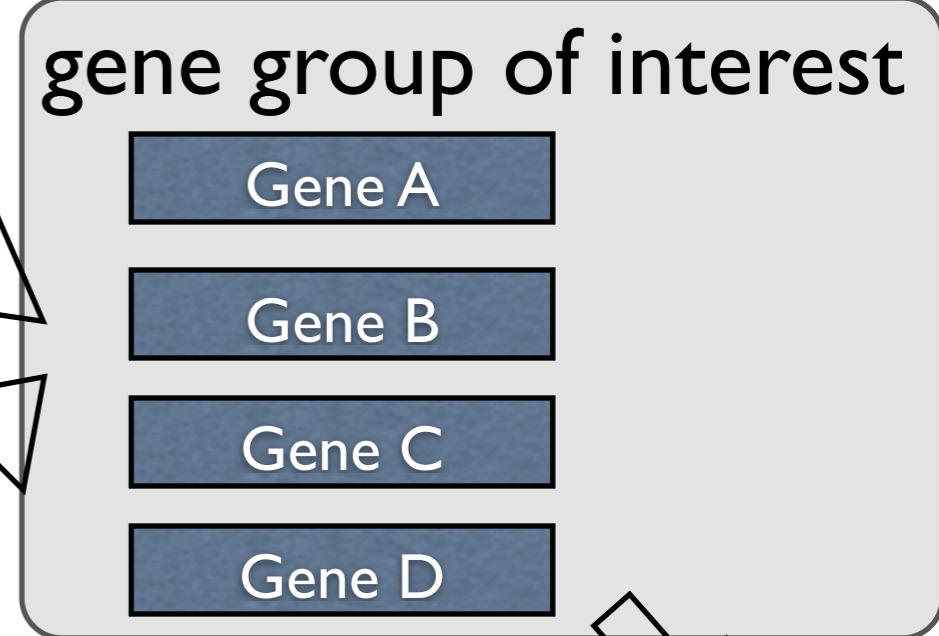
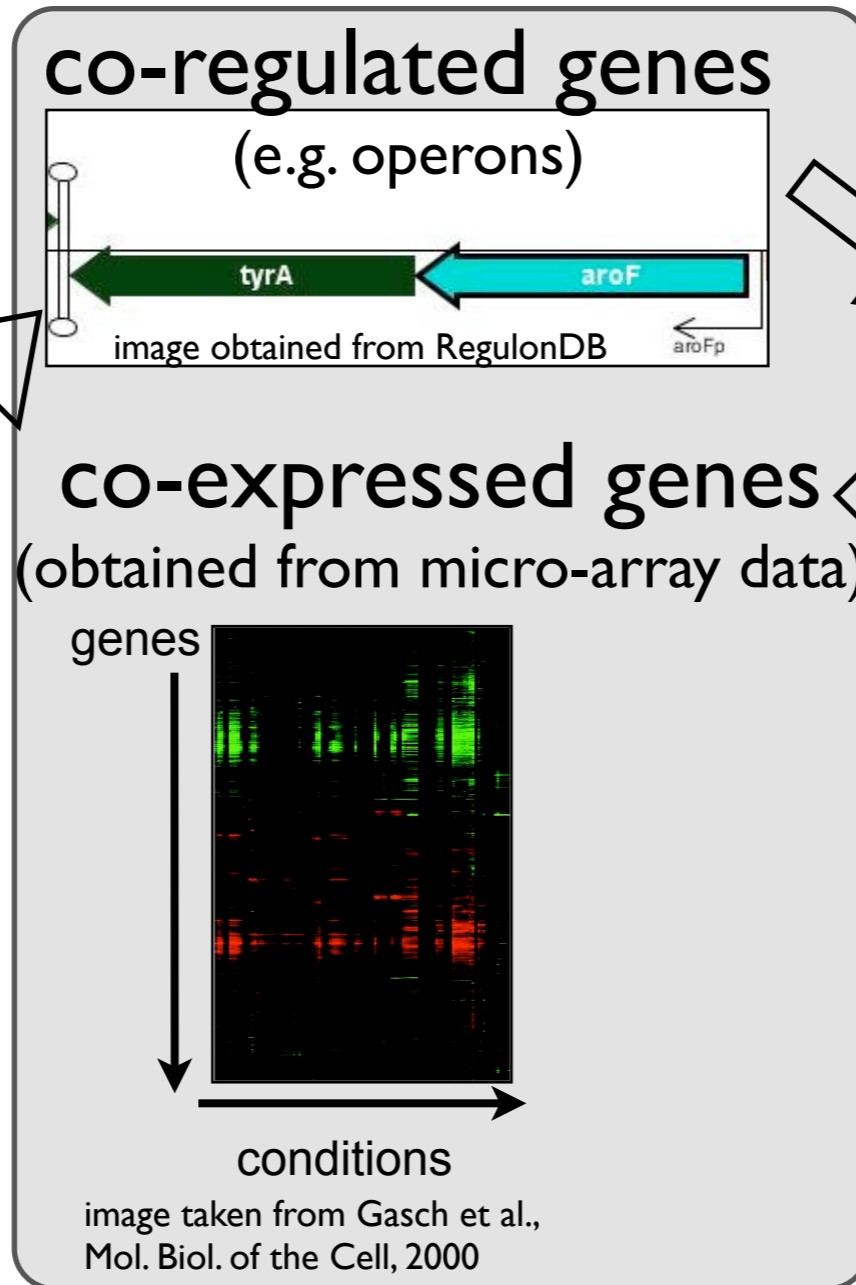
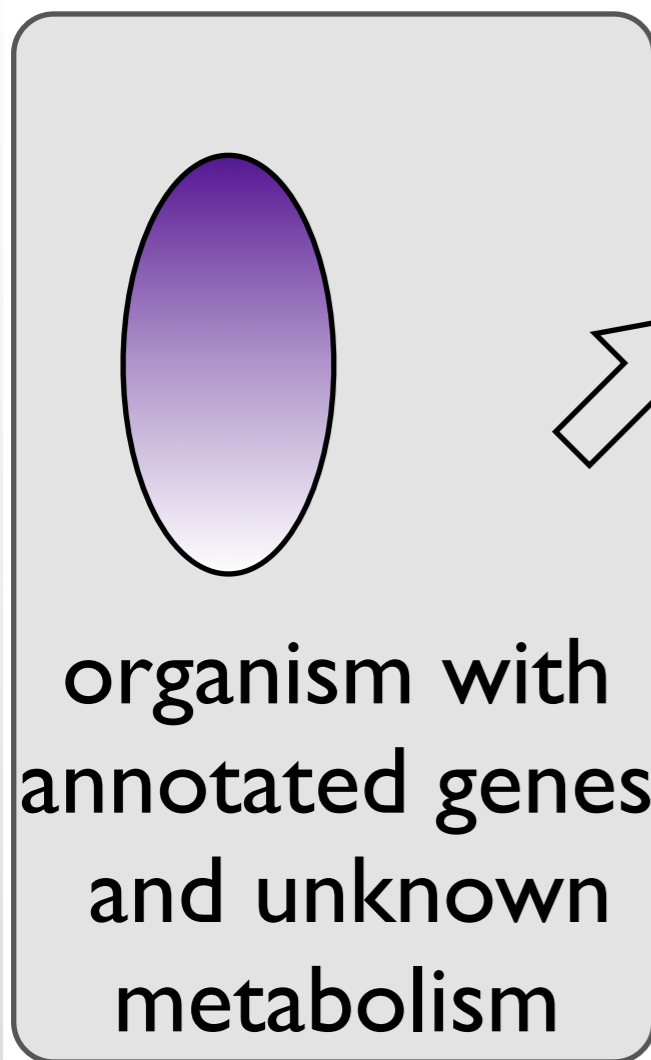
gene group of interest

- Gene A
- Gene B
- Gene C
- Gene D

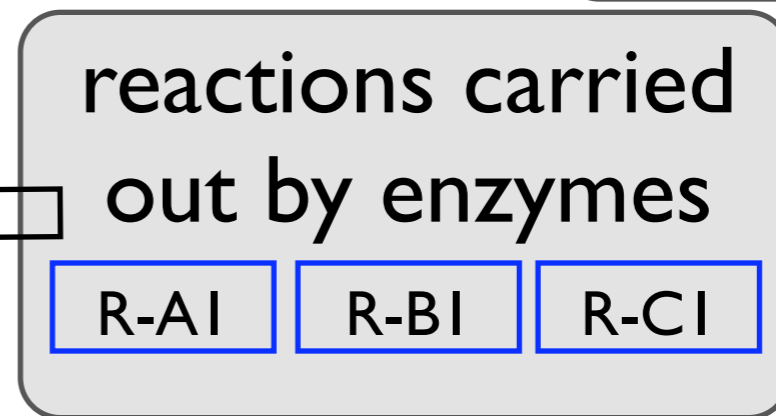
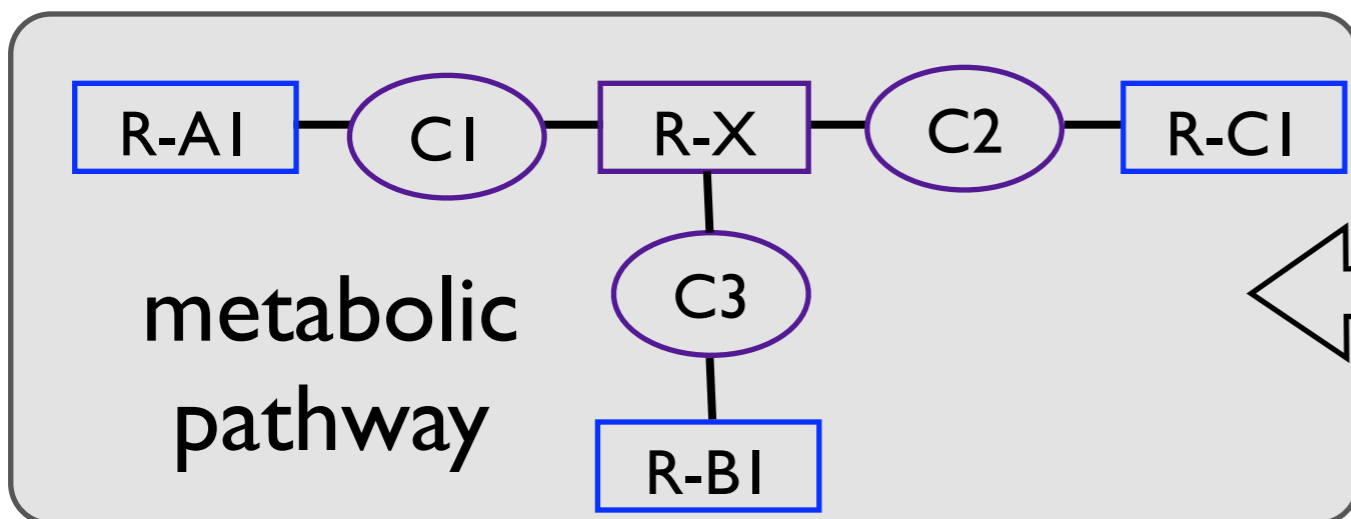
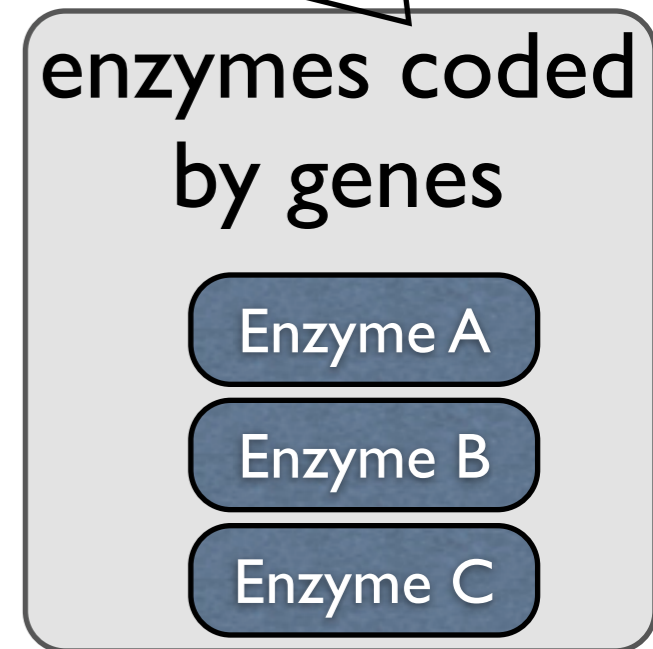
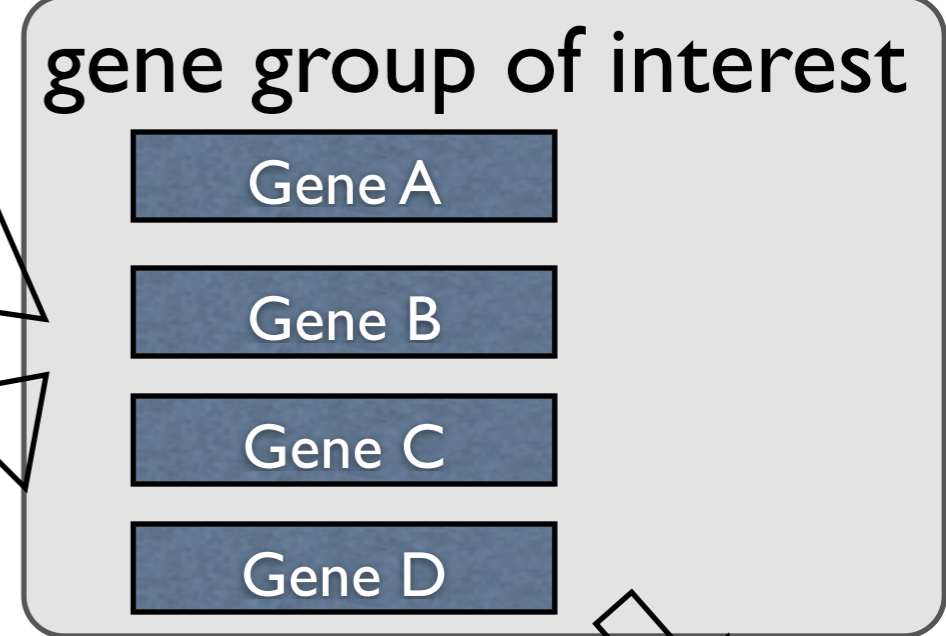
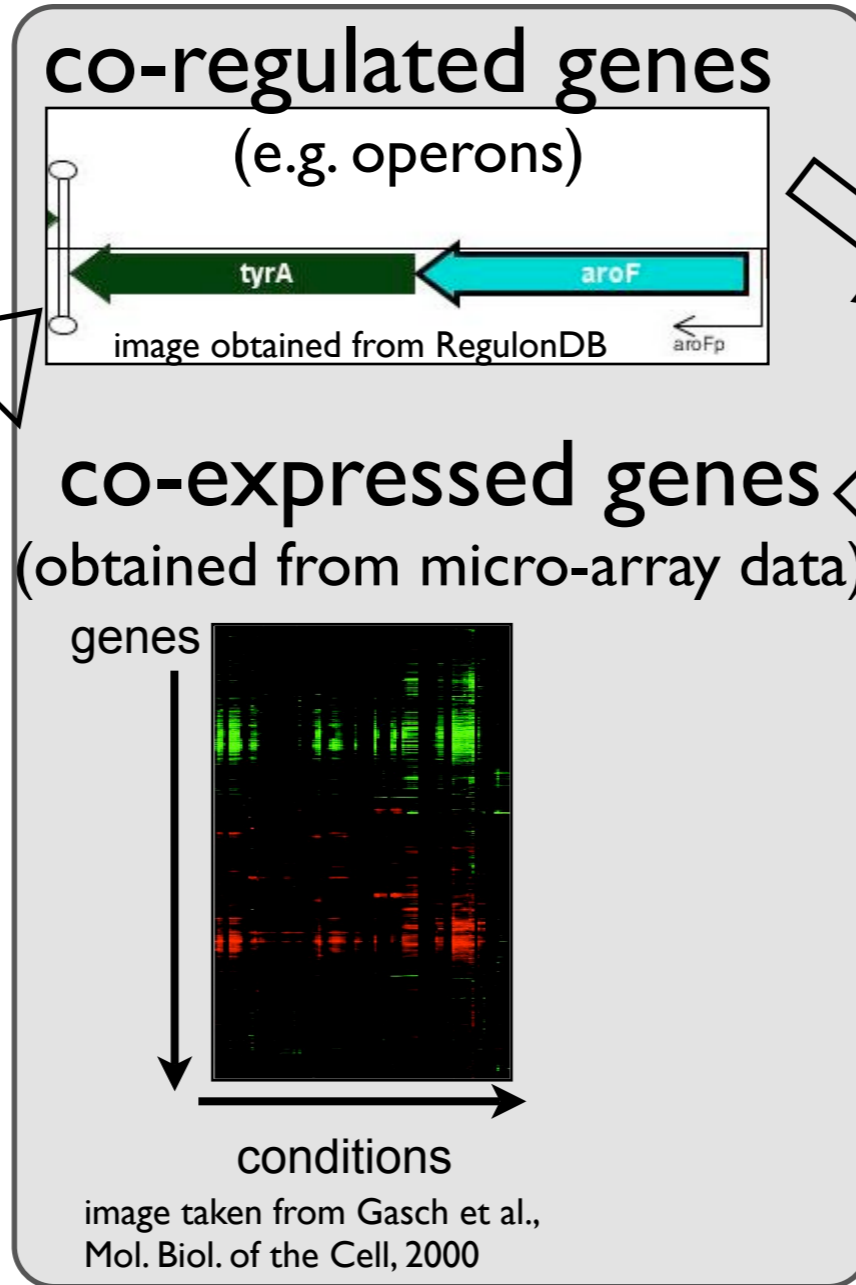
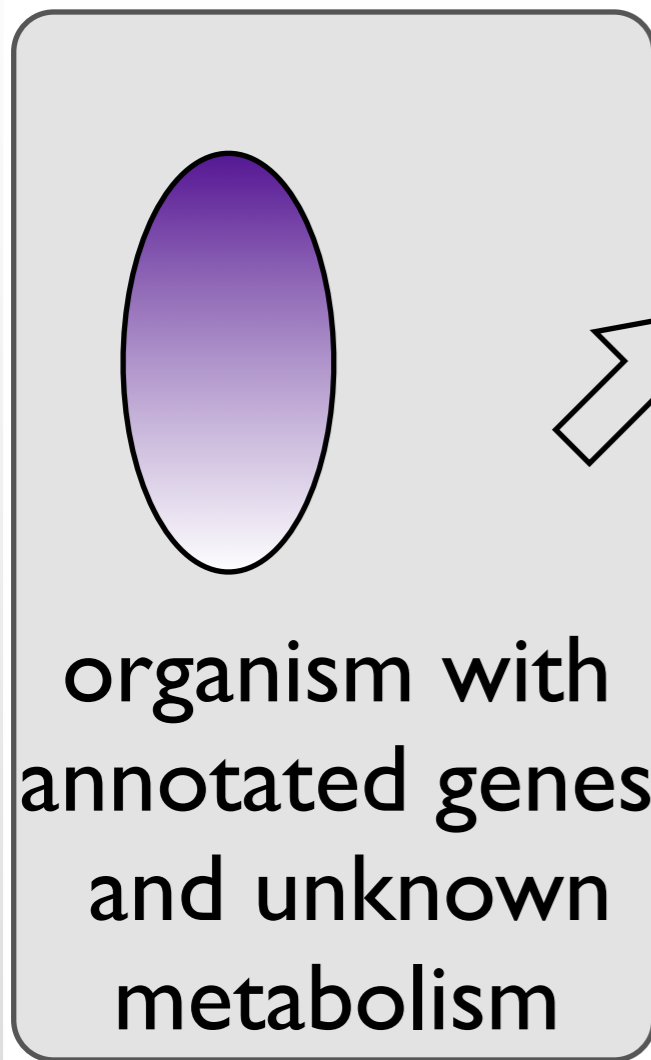
enzymes coded by genes

- Enzyme A
- Enzyme B
- Enzyme C

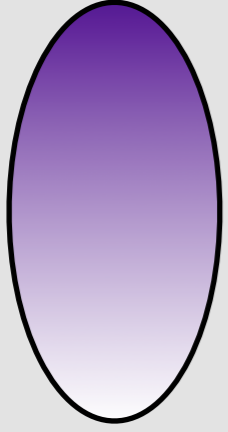
Biological question



Biological question



Biological question



organism with annotated genes and unknown metabolism

co-regulated genes (e.g. operons)

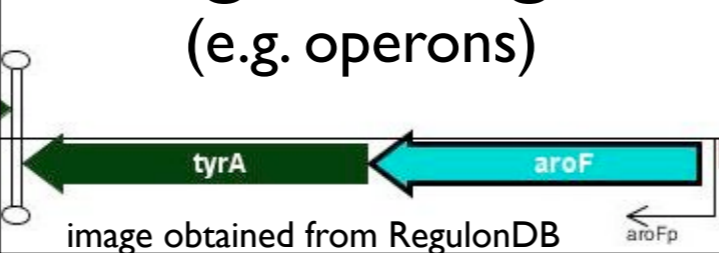


image obtained from RegulonDB

co-expressed genes (obtained from micro-array data)



genes

conditions

image taken from Gasch et al., Mol. Biol. of the Cell, 2000

gene group of interest

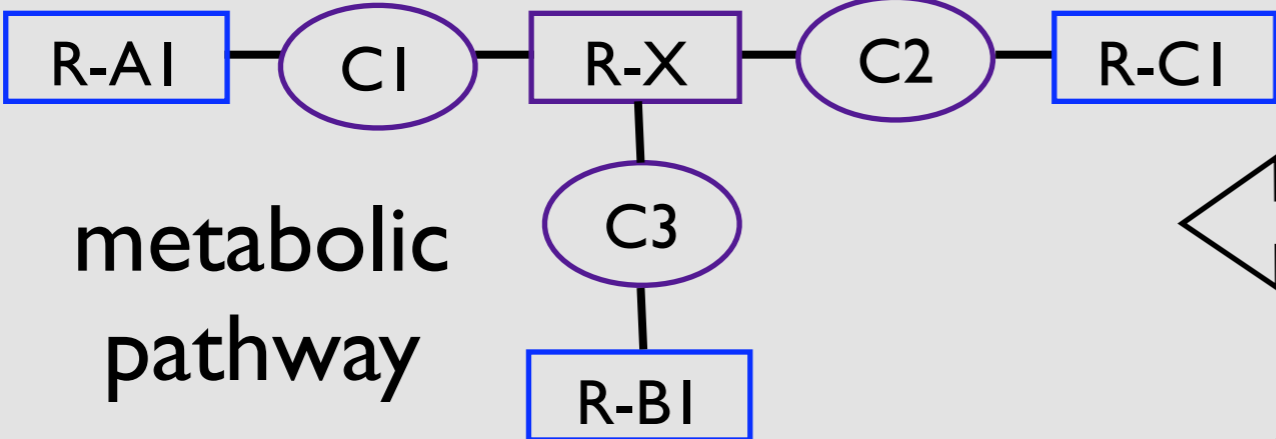
- Gene A
- Gene B
- Gene C
- Gene D

enzymes coded by genes

- Enzyme A
- Enzyme B
- Enzyme C

In which metabolic pathway(s) participate the enzymes coded by genes assumed to be functionally related?

metabolic pathway



```
graph LR; R-AI --> C1((C1)); C1 --> R-X[R-X]; R-X --> C2((C2)); C2 --> R-CI[R-CI]; R-X --> C3((C3)); C3 --> R-BI[R-BI]
```

reactions carried out by enzymes

- R-AI
- R-BI
- R-CI

Introduction

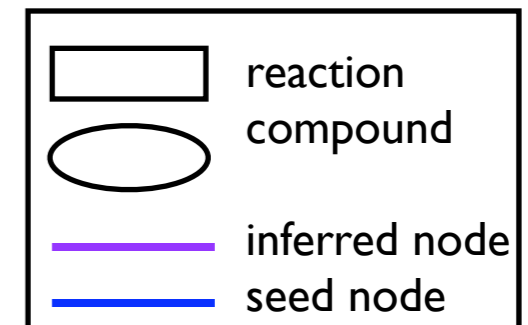
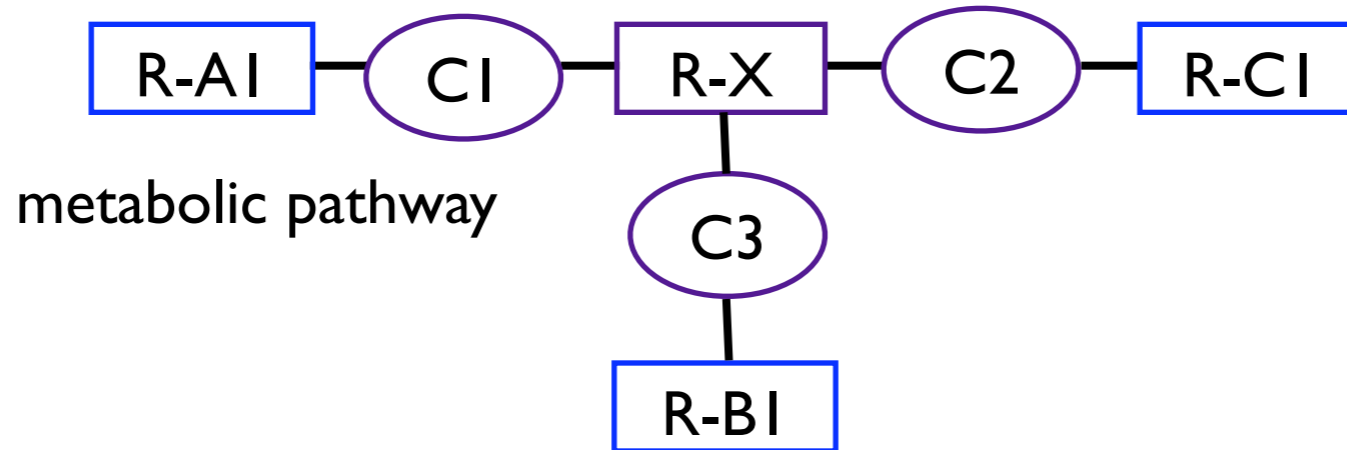
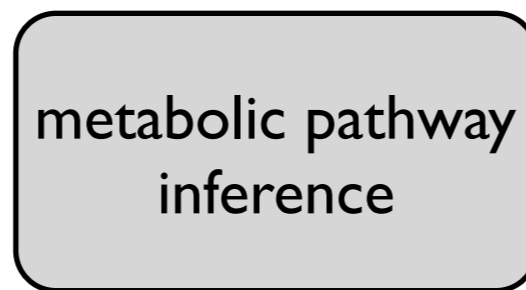
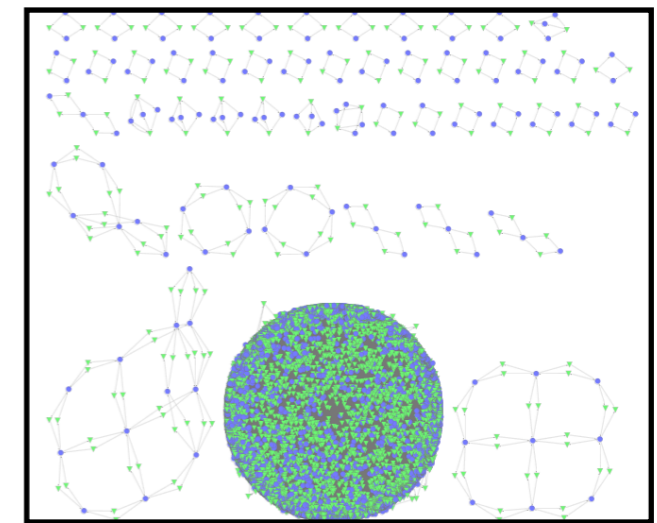
Metabolic pathway inference - principle

given a set of seed reactions, find meaningful pathways connecting them in a metabolic graph

seed reactions



metabolic graph (containing all known reactions and compounds)

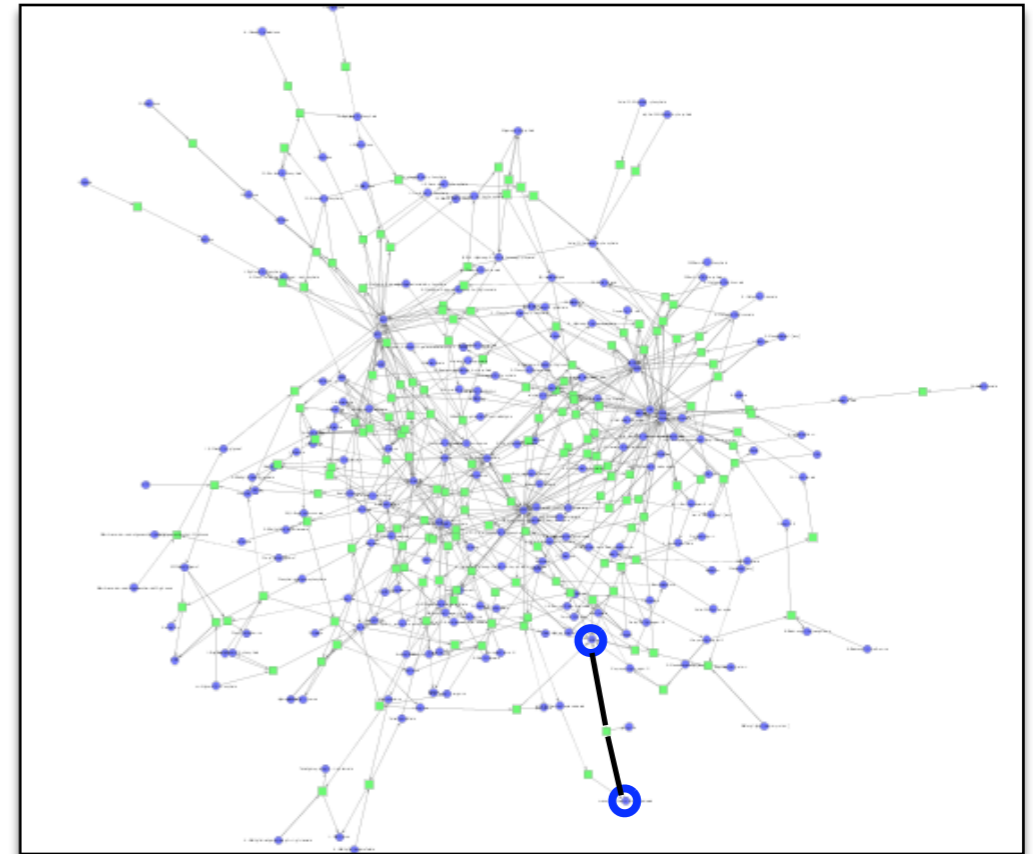


Pathway inference methods

How can we extract subgraphs
(pathways) from metabolic graphs?

Two-end path finding - principle

- idea: infer pathway given two **seed nodes** using a path finding algorithm (*k*-shortest paths algorithm)
- problem: highly connected compounds (such as H₂O and ATP) are preferentially traversed

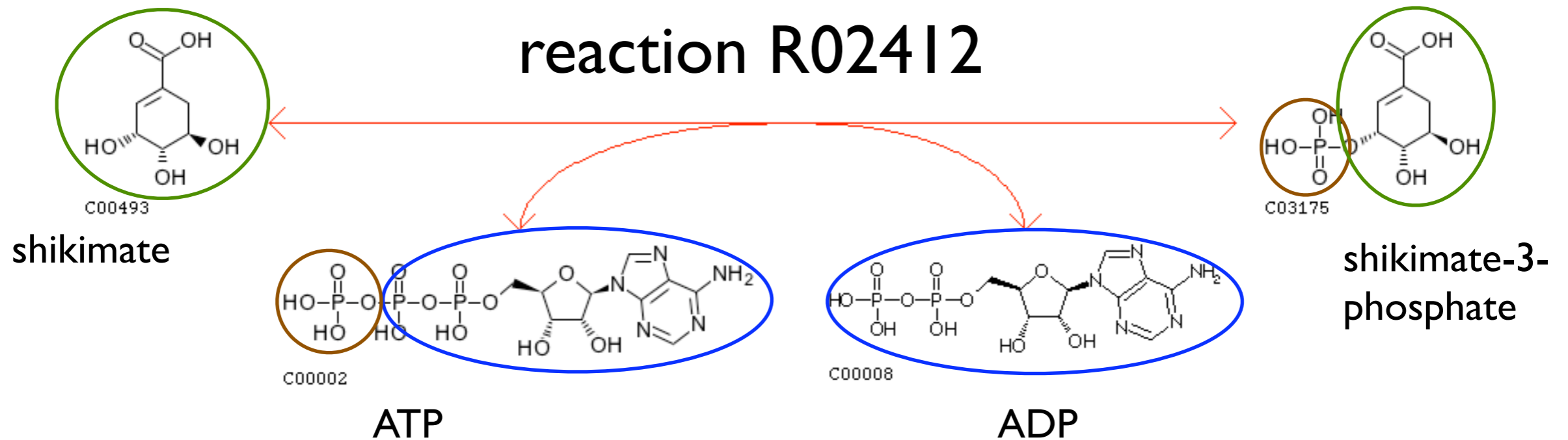


Two-end path finding - reaction traversal

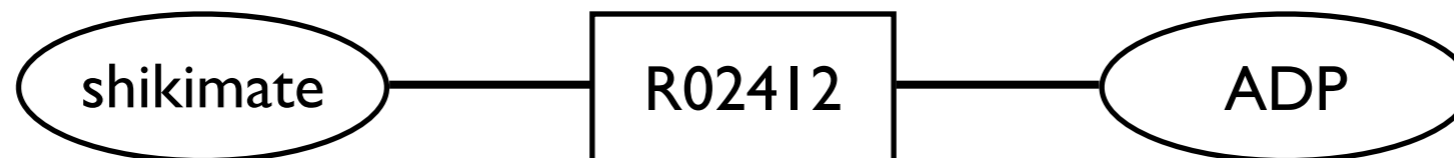
How to traverse a reaction?

substrates

products



traversal from main to side compound: biochemically irrelevant



Two-end path finding - evaluation

Graph type	Average geometric accuracy
weighted KEGG graph incorporating main-side compound annotation (weighted KEGG RPAIR graph)	83%
unweighted KEGG graph incorporating main-side compound annotation (unweighted KEGG RPAIR graph)	72%
weighted KEGG graph	73%
filtered KEGG graph (hub compounds removed)	57%
unweighted KEGG graph	16%

evaluation on 55 linear pathways from three organisms (*E. coli*, *S. cerevisiae*, *H. sapiens*)

D. Croes, F. Couche, S. Wodak and J. van Helden (2006). "Inferring Meaningful Pathways in Weighted Metabolic Networks." *J. Mol. Biol.* 356: 222-236.

D. Croes, F. Couche, S. Wodak and J. van Helden (2005). "Metabolic PathFinding: inferring relevant pathways in biochemical networks." *Nucleic Acids Research* 33: W326-W330.

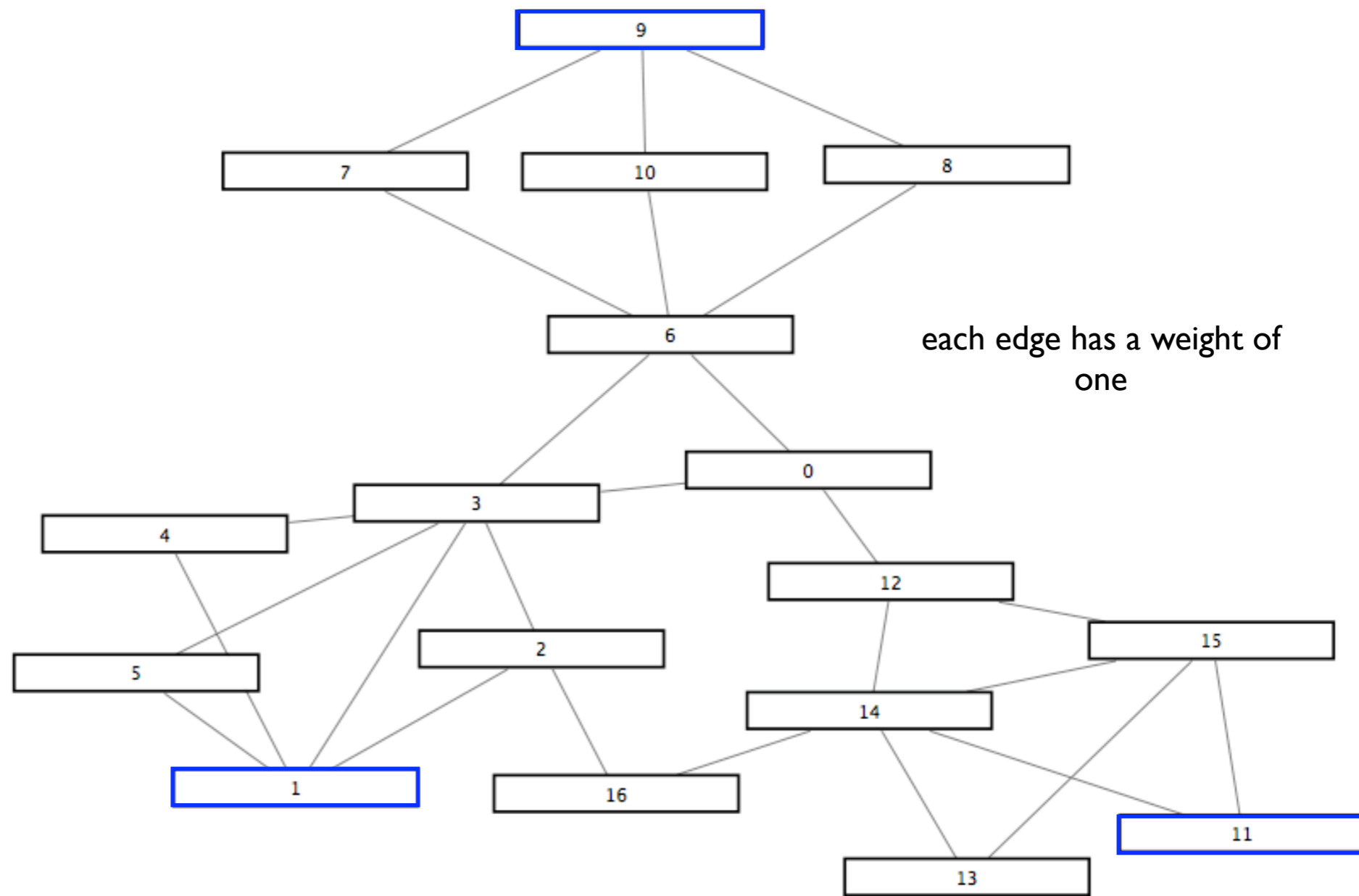
Kotera, M., Hattori, M., Oh, M.-A., Yamamoto, R., Komeno, T., Yabuzaki, J., Tonomura, K., Goto, S., and Kanehisa, M. (2004). "RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions" *Genome Informatics* 15.

K. Faust, D. Croes and J. van Helden (2008). "Metabolic path finding using RPAIR annotation." Submitted.

Multiple-end pathway inference

Pairwise k -shortest paths - principle

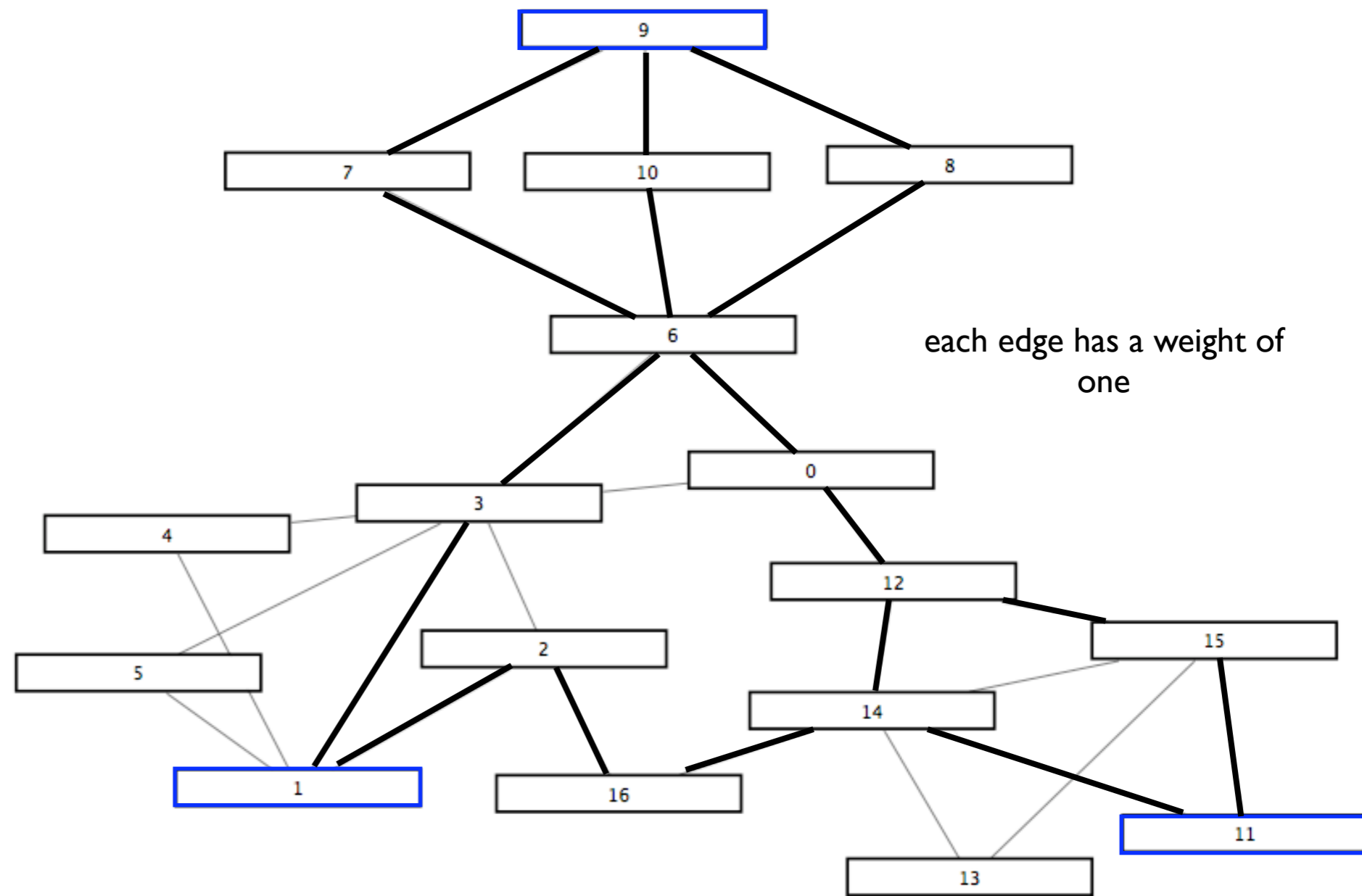
- extend two-end path finding to **multiple seeds** pathway inference by calling a k -shortest paths algorithm (REA) repetitively



Multiple-end pathway inference

Pairwise k -shortest paths - paths computation

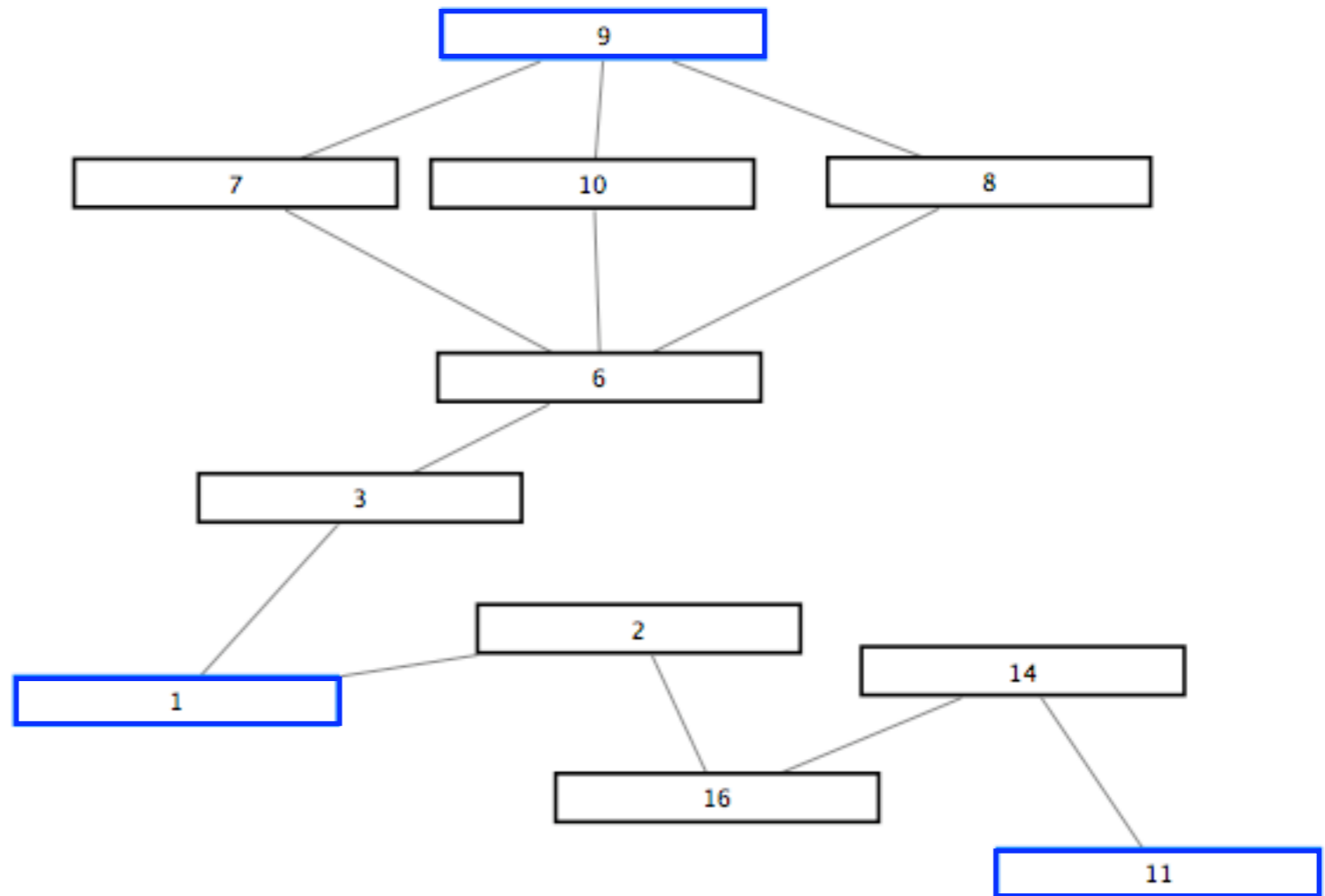
- for each seed node pair, obtain **all lightest paths** with a k -shortest paths algorithm



Multiple-end pathway inference

Pairwise k -shortest paths - subgraph extraction

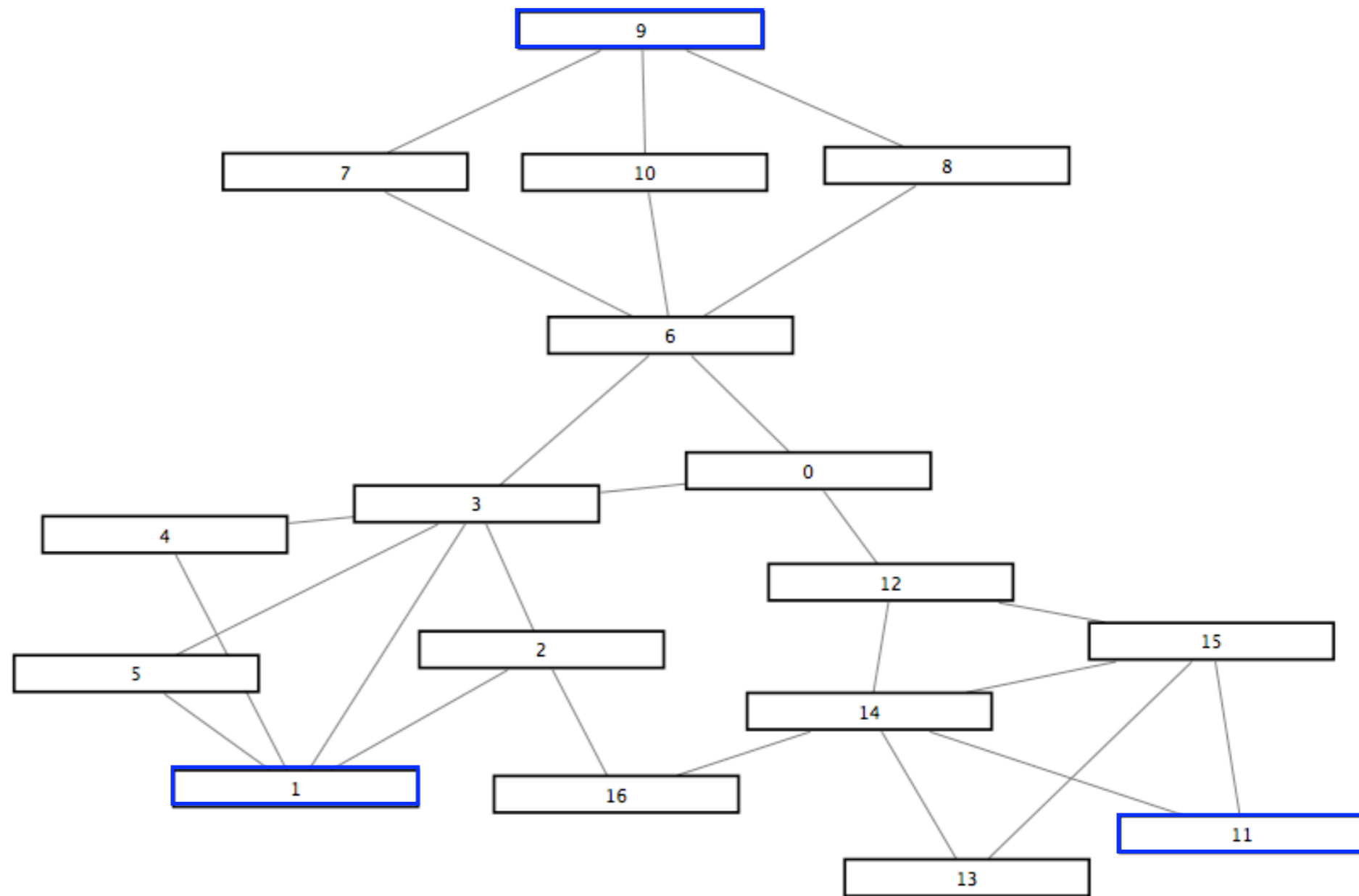
- merge lightest paths
in the order of their
weight until either all
seed nodes are
connected or all
lightest paths are
merged



Multiple-end pathway inference

kWalks algorithm - principle

- idea: some edges and nodes in a graph are more relevant than others to connect given **seed nodes**

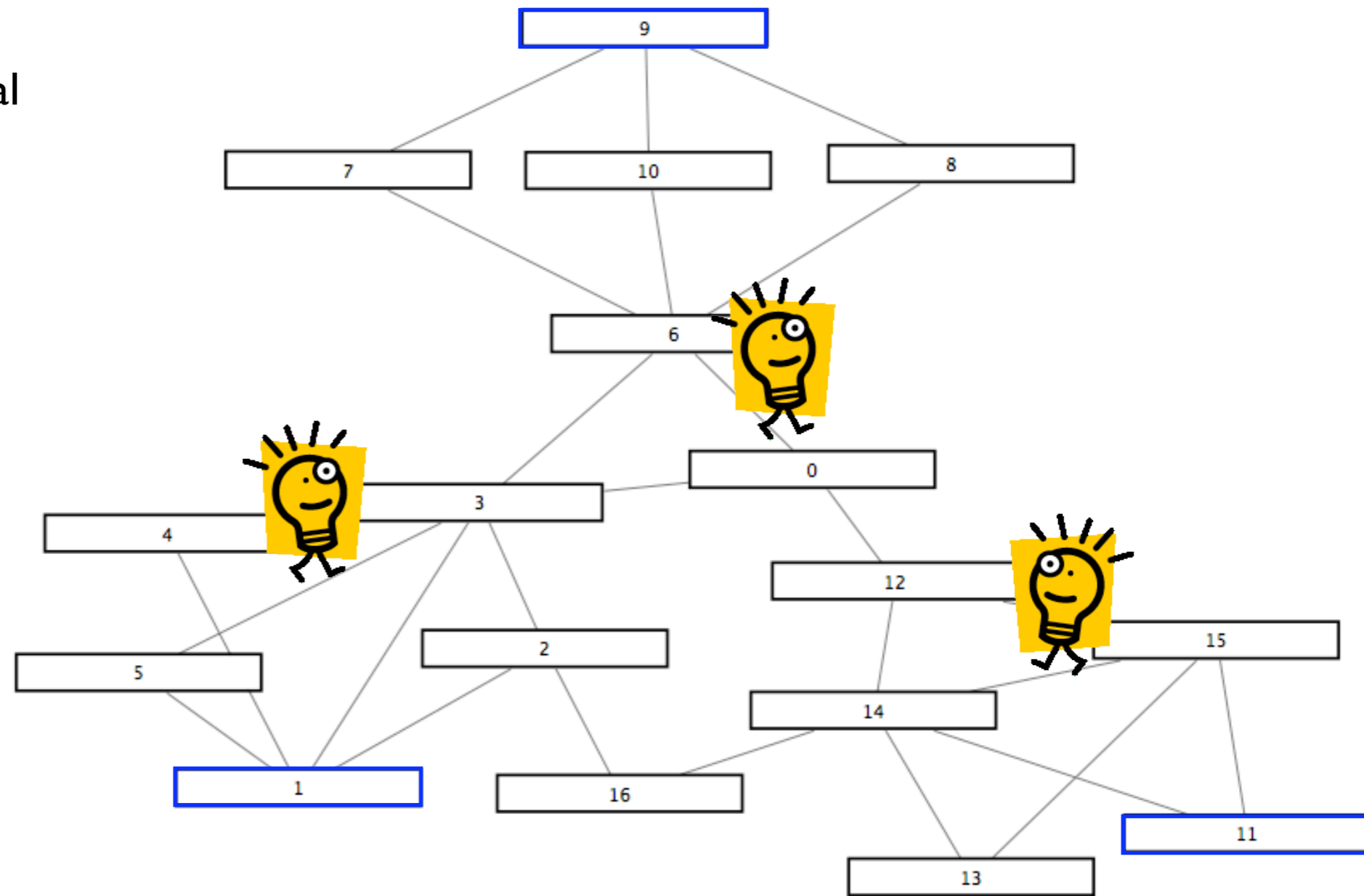


P. Dupont, J. Callut, G. Doms, J.-N. Monette and Y. Deville (2006-2007). "Relevant subgraph extraction from random walks in a graph." Research Report UCL/FSA/INGI RR 2006-07, November 2006.

Multiple-end pathway inference

kWalks algorithm - edge relevance computation

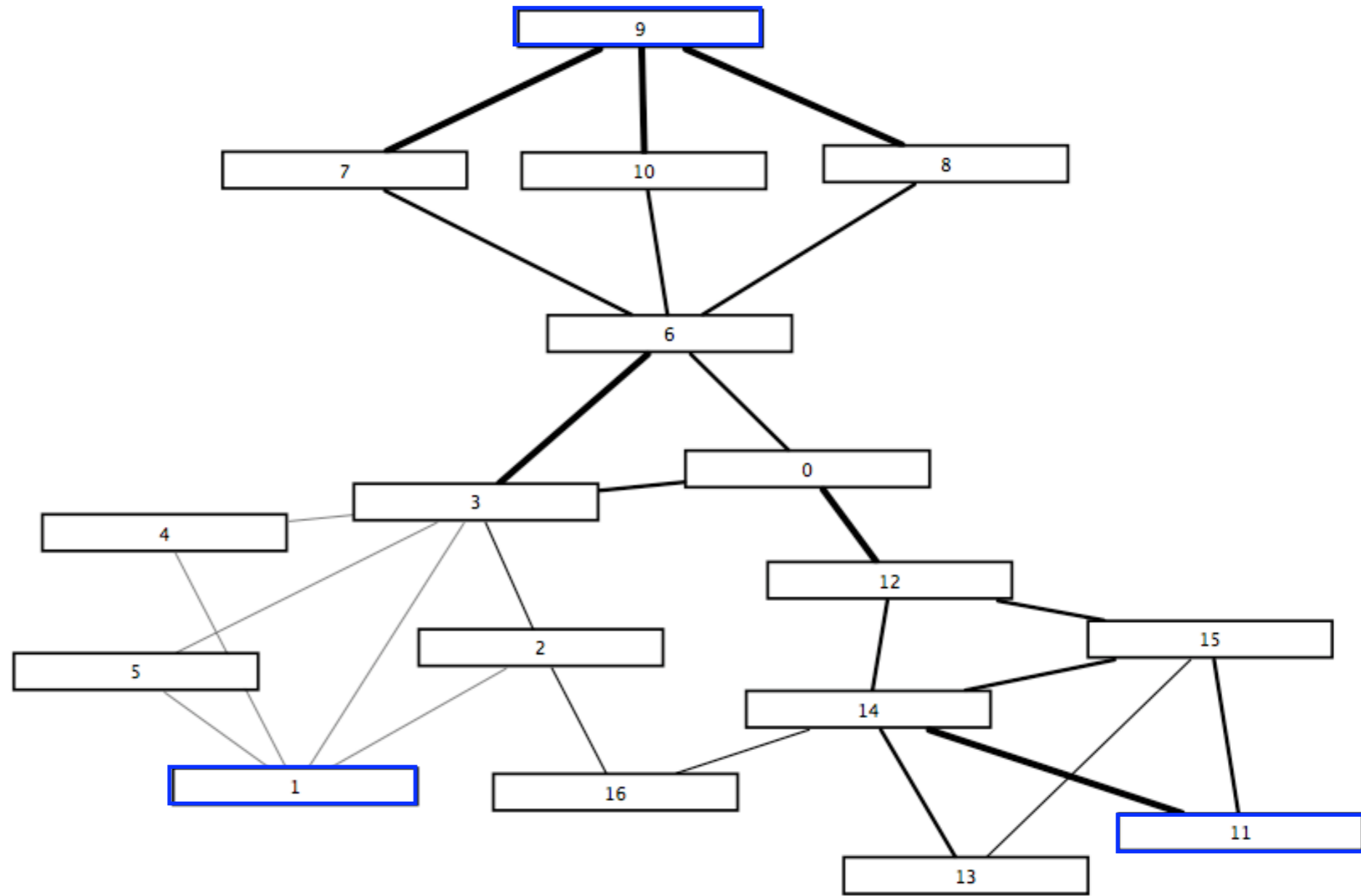
- edge or node relevance: proportional to the expected number of times it is visited by **random walkers**, each starting from one of the **seed nodes**



Multiple-end pathway inference

kWalks algorithm - output

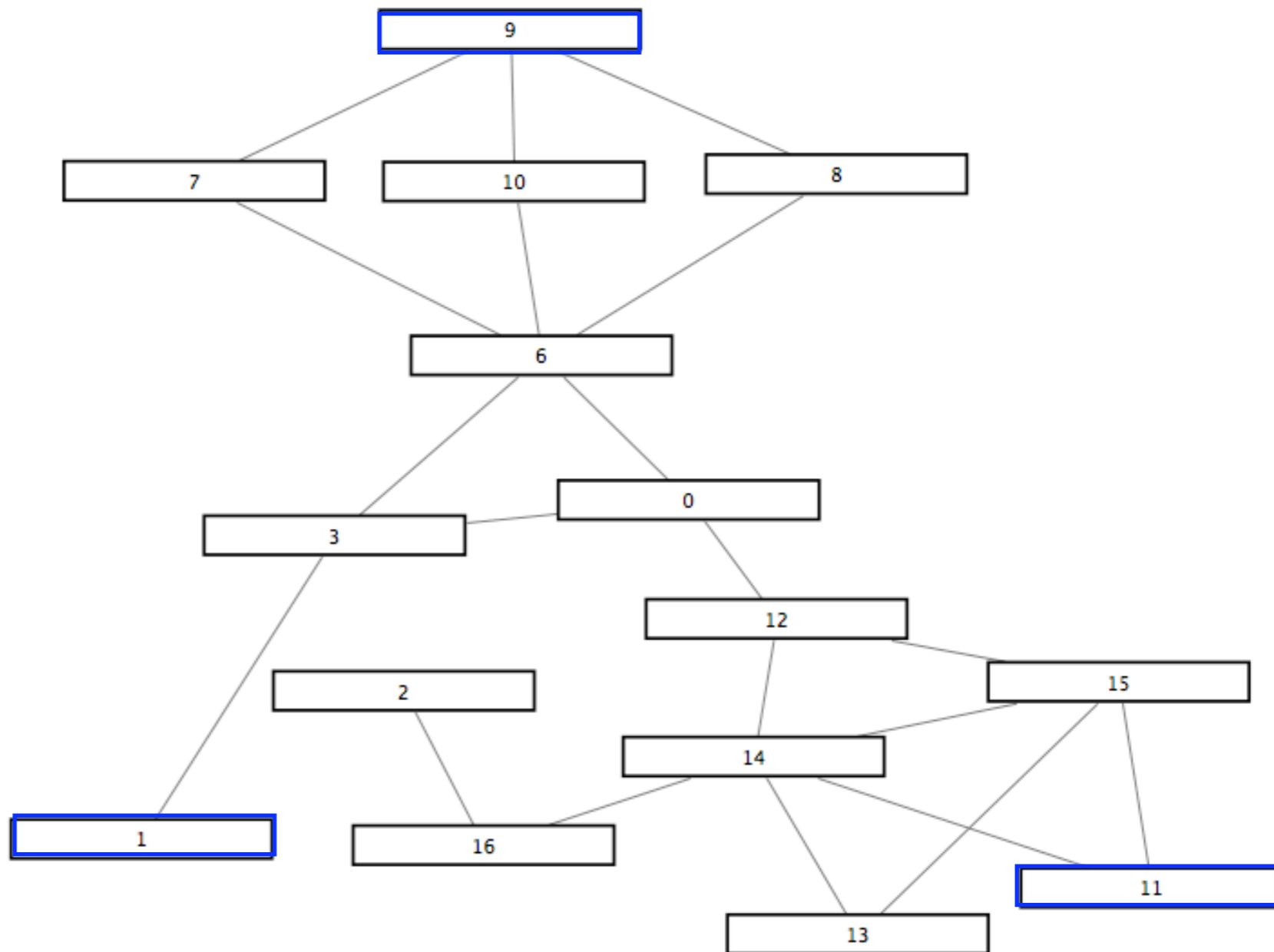
- list of edge
and node
relevances



Multiple-end pathway inference

kWalks algorithm - subgraph extraction

- add edges and their adjacent nodes in the order of their relevance to the seed nodes until seed nodes are connected or no more edges can be added

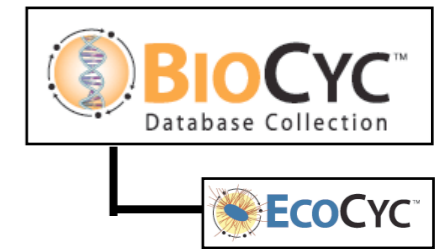
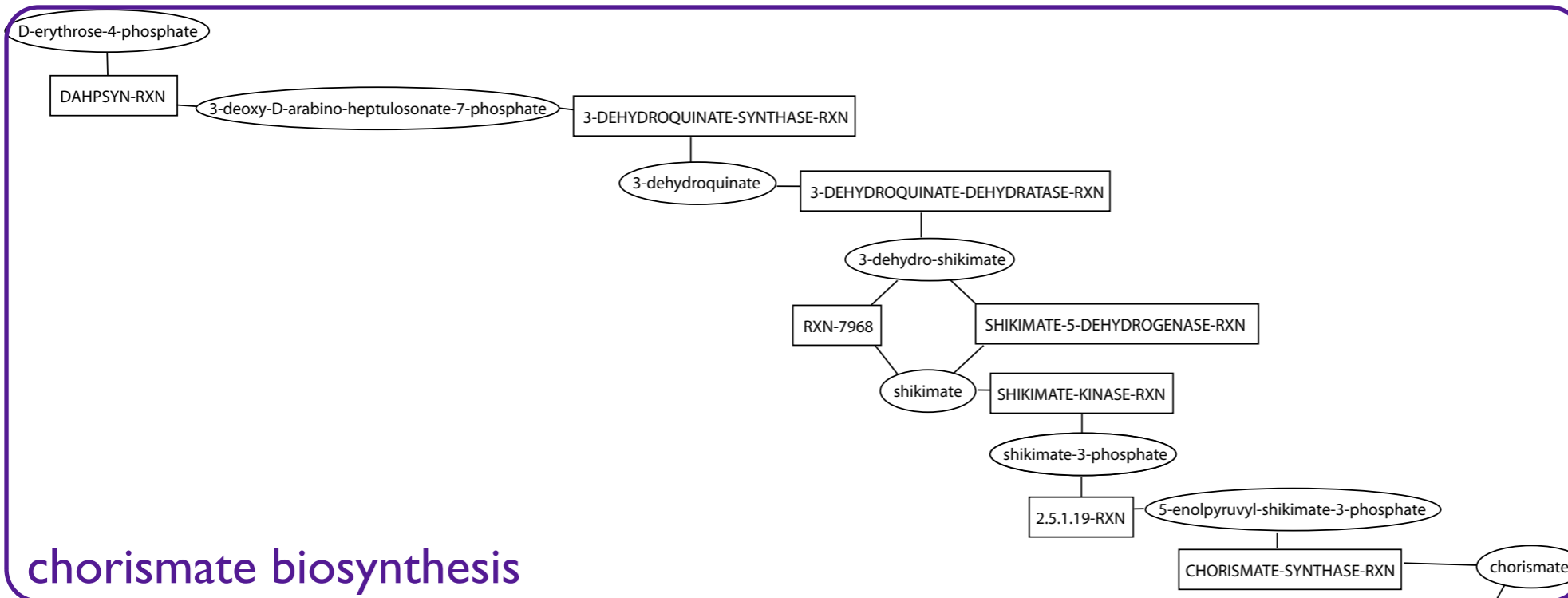


Pathway inference evaluation

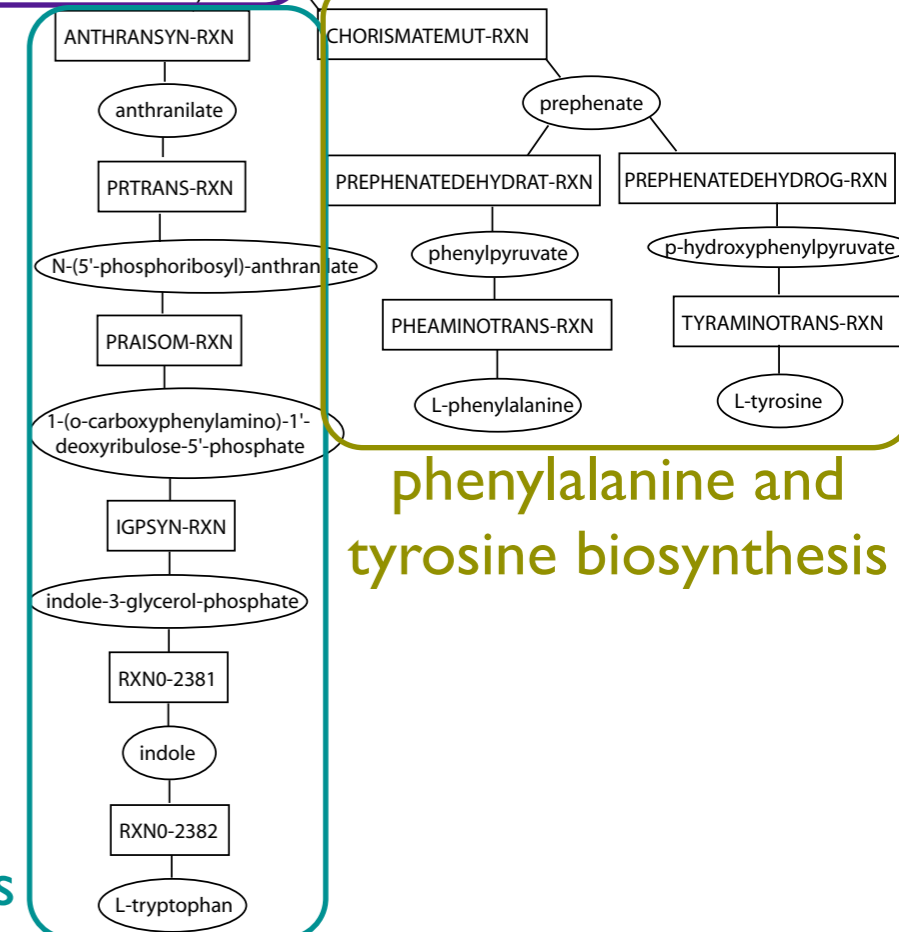
How accurately can these algorithms infer known pathways from metabolic graphs?

Pathway inference evaluation - example

Aromatic amino acid biosynthesis (*E. coli*)



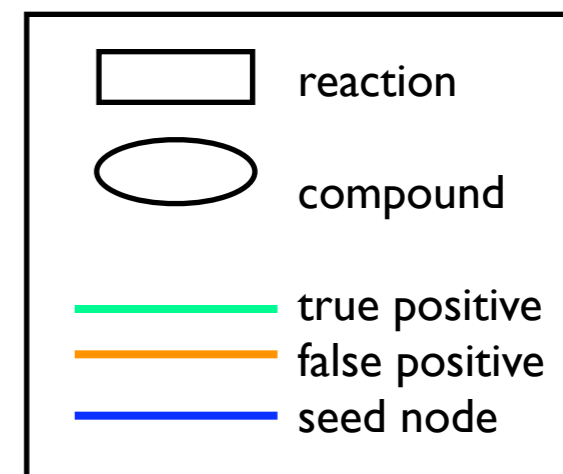
- reference pathway consisting of:
- 15 compounds (without terminal compounds)
 - 19 reactions
 - 3 branches (leading to the 3 aromatic amino acids tryptophan, phenylalanine and tyrosine)



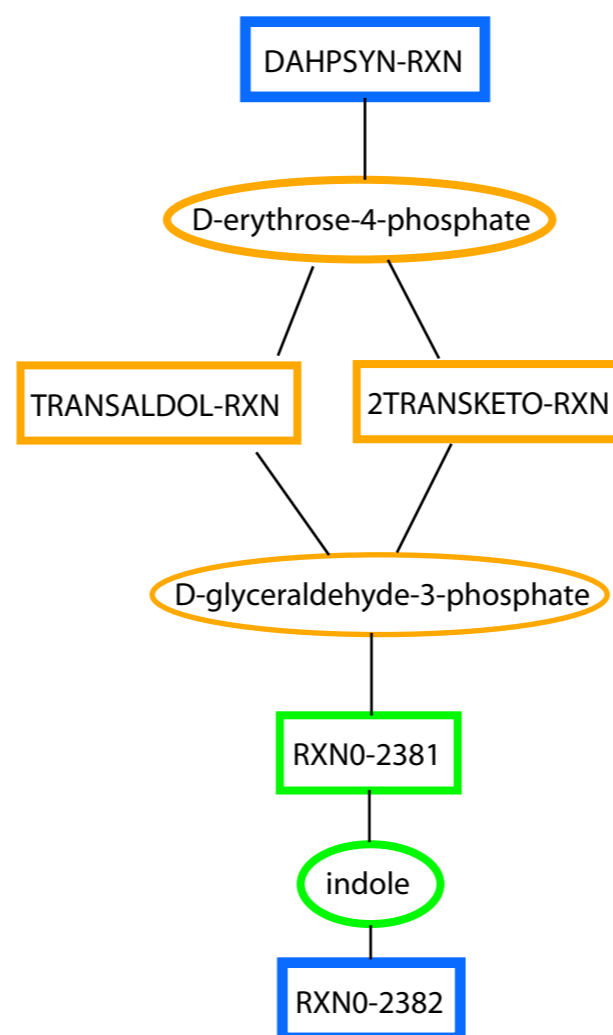
tryptophan biosynthesis

Pathway inference evaluation - example

Pathway inferred with 2 seed reactions



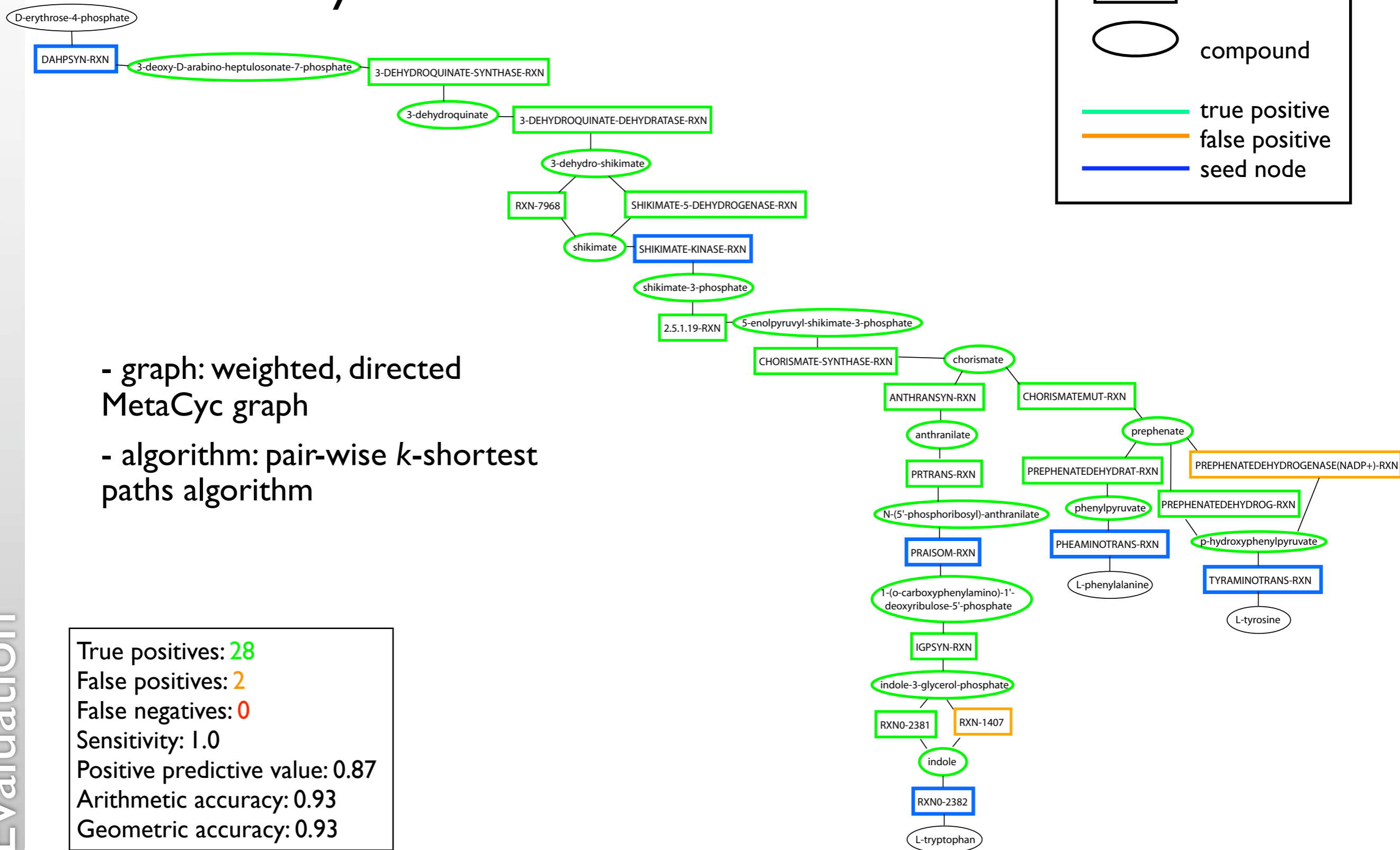
- graph: weighted, directed
MetaCyc graph
- algorithm: pair-wise k -shortest
paths algorithm



True positives: 2
False positives: 4
False negatives: 32
Sensitivity: 0.06
Positive predictive value: 0.33
Arithmetic accuracy: 0.2
Geometric accuracy: 0.02

Pathway inference evaluation - example

Pathway inferred with 6 seed reactions



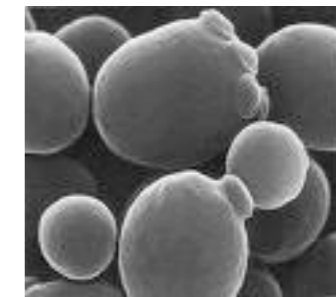
- graph: weighted, directed MetaCyc graph
- algorithm: pair-wise *k*-shortest paths algorithm

True positives: 28
 False positives: 2
 False negatives: 0
 Sensitivity: 1.0
 Positive predictive value: 0.87
 Arithmetic accuracy: 0.93
 Geometric accuracy: 0.93

Pathway inference evaluation in MetaCyc

Reference pathways

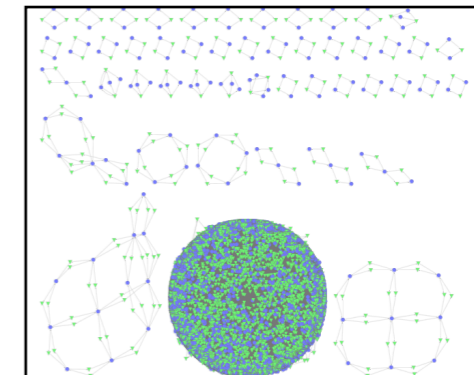
- 71 pathways taken from the *Saccharomyces cerevisiae* pathways annotated in MetaCyc (curated tier of BioCyc)
- minimal pathway size: 5 nodes
- average node number: 13
- 34 branched and 17 cyclic pathways



Saccharomyces cerevisiae, taken from <http://www.bath.ac.uk/bio-sci/wheals2.htm>

Metabolic graph

- MetaCyc Release 11.0 (all small molecule compounds and their reactions)
- 4,891 compound nodes and 5,358 reaction nodes

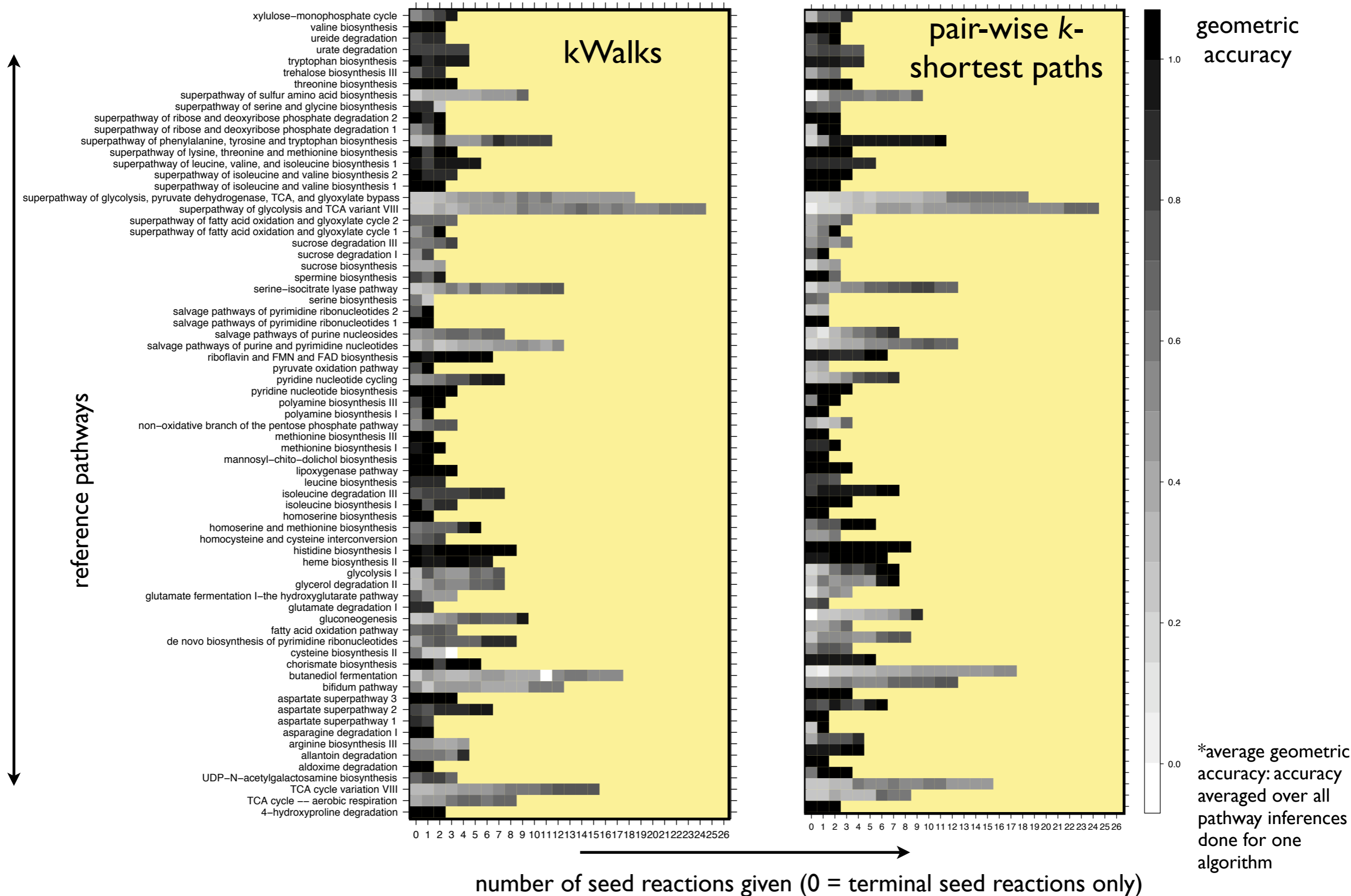


MetaCyc metabolic graph displayed in Cytoscape

Evaluation procedure

- for each reference pathway, do inference with terminal reactions of the reference pathway as seed nodes
- repeat inference by adding one additional randomly chosen reaction at each step to the seed reaction set

Evaluation in weighted MetaCyc graph



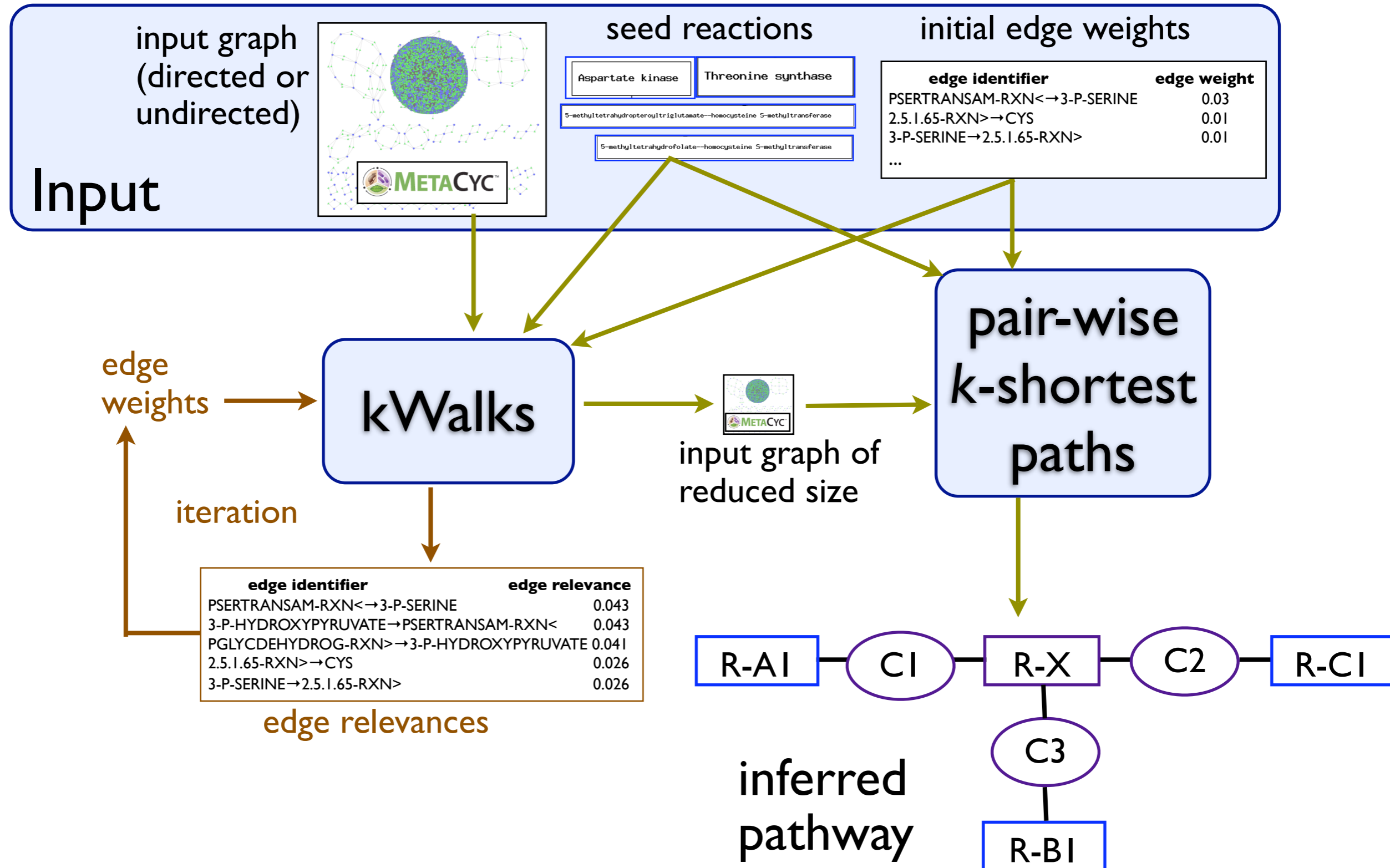
average geometric accuracy*: ~60%

average geometric accuracy*: ~68%

Pathway inference evaluation - results

- kWalks is quick (order of seconds) and has high sensitivity, but lower positive predictive value than pair-wise k -shortest paths
- pair-wise k -shortest paths: high geometric accuracy, but is too slow (runtime increases quadratically with seed node number!)

Parameter tuning



Parameter tuning - Parameters and their values

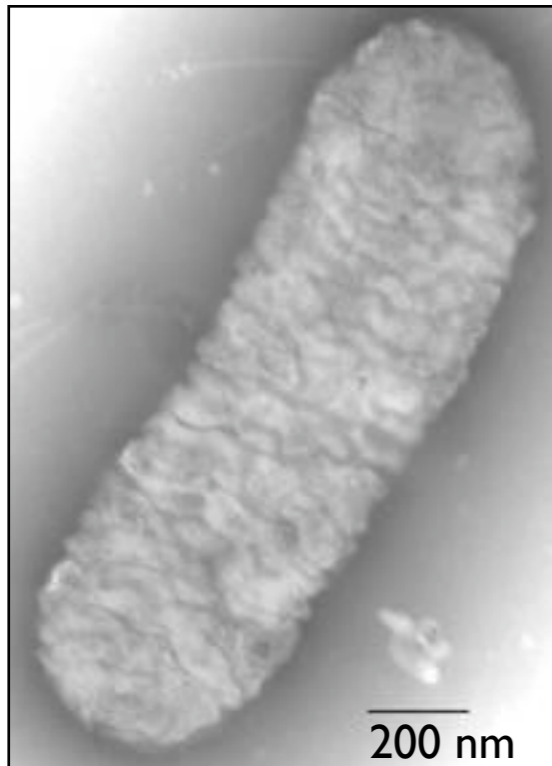
Parameter	Values
Algorithm	KWalks, pair-wise k -shortest paths, hybrid (combination of kWalks and pair-wise k -shortest paths)
Input edge weights	Unit (all weights set to 1), compound degree (reactions: weight of 1, compounds: node degree as weight), inflation of weights (weight to the power of positive integer)
KWalks iteration number	1, 3 and 6
Hybrid: use of kWalks edge relevances as weights in pair-wise k -shortest paths	True/False
Graph directionality	Directed (including direct and reverse direction for each reaction)/undirected
Hybrid: size of subgraph extracted by kWalks	0.1% to 10% of input graph edge number

Parameter tuning - results

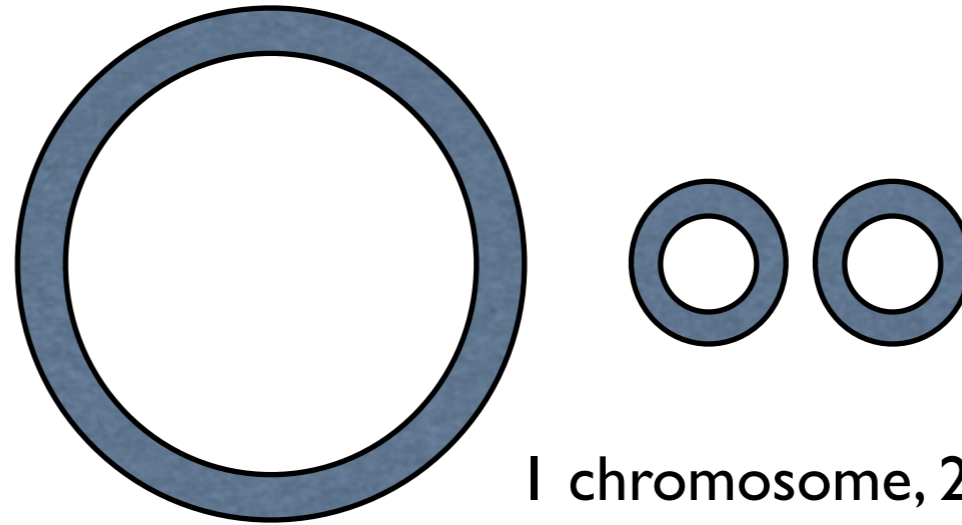
- positive predictive value of kWalks can be increased by iteration and combination with pair-wise k -shortest paths (hybrid approach)
- with optimal parameter values set, kWalks and pair-wise k -shortest paths reach similar average geometric accuracies ($\sim 68\%$)
- the hybrid algorithm (with optimal fixed subgraph size) yields an average geometric accuracy of 72%

Analysis of *R. metallidurans* operons

Ralstonia metallidurans CH34 (*Cupriavidus metallidurans* CH34)



Ralstonia metallidurans (*Cupriavidus metallidurans*) CH34
© Groupe Toxicologie humaine et environnementale, Laboratoire Pierre Süe, UMR 9956 CNRS/CEA Saclay/Centre commun de microscopie électronique d'Orsay



1 chromosome, 2 mega plasmids (pMol30 and pMol28), overall size: ~6 Mb

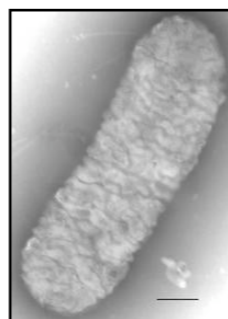
- gram-negative bacterium
- resistance to heavy metals (zinc, nickel, cadmium, cobalt, copper, ...)
- metabolism only reconstructed by automatic procedures (e.g. PathoLogic)
- 6,176 protein-coding genes, of those 832 enzymes (source: BioCyc)

Analysis of *R. metallidurans* operons



Seed reactions

- 4,060 operons predicted in *Ralstonia metallidurans*¹
- KEGG provides *R. metallidurans* gene-reaction mappings
- 294 operons could be associated to more than one reaction



Metabolic graph

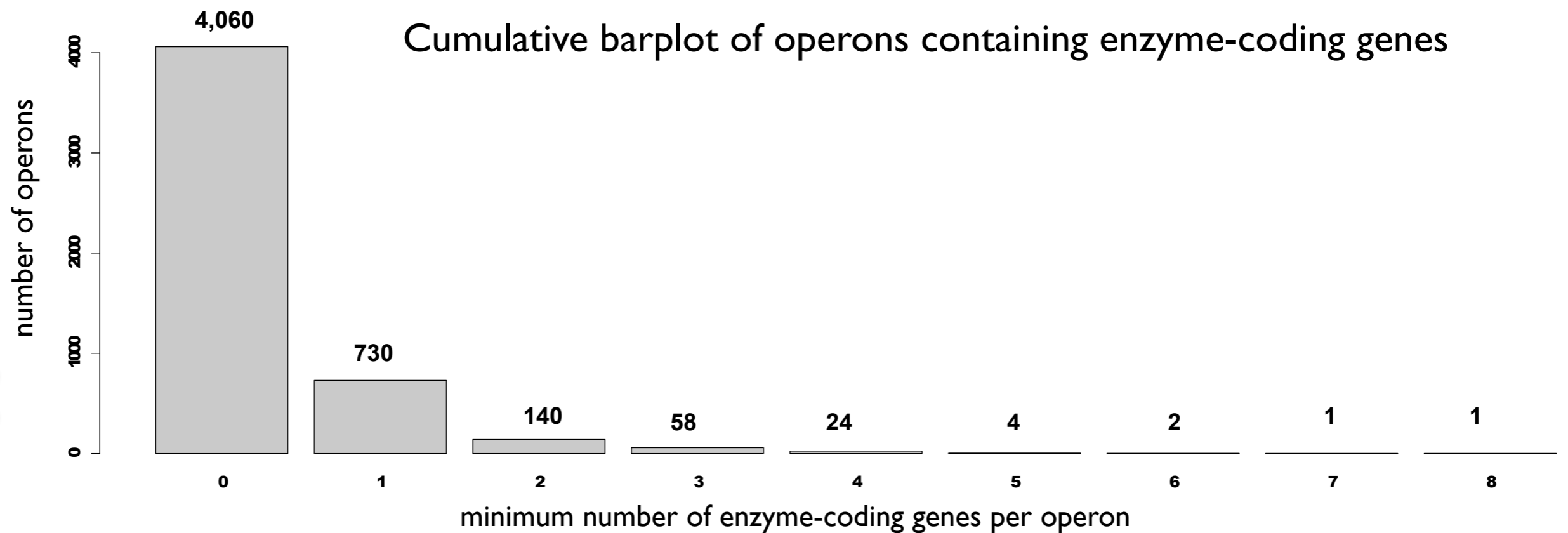
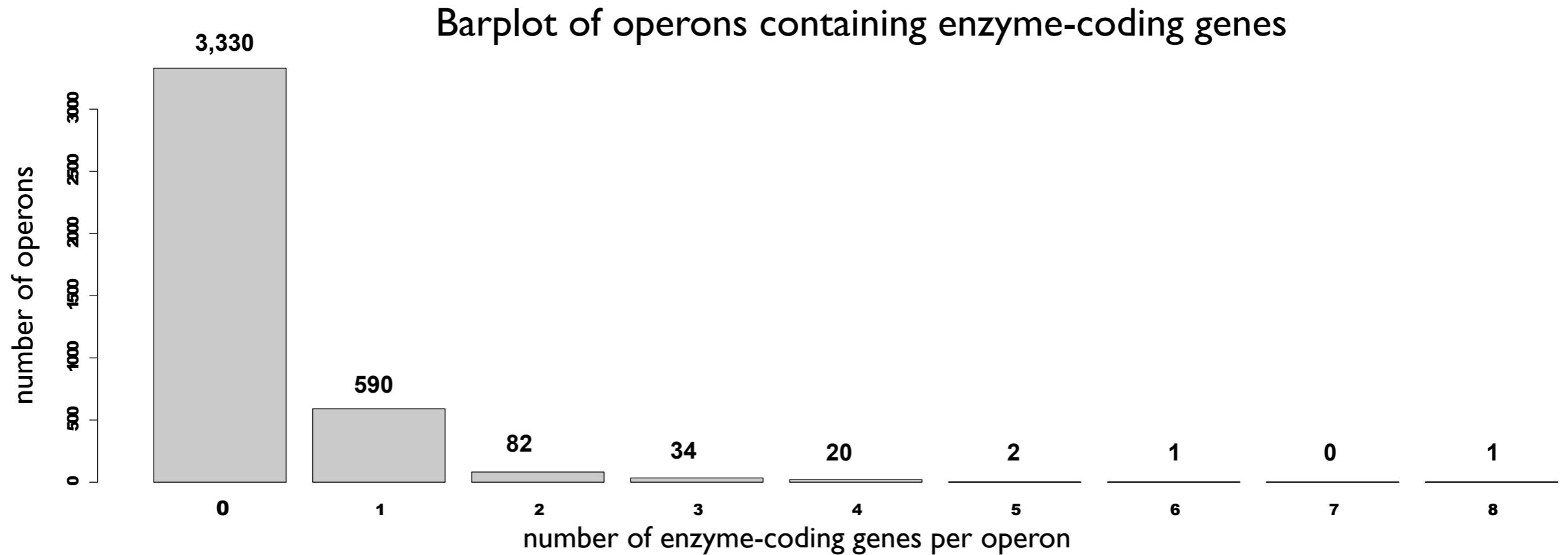
- weighted KEGG RPAIR graph incorporating main/side compound annotation (KEGG version 41.0)
- graph contains all compounds and reactant pairs in KEGG RPAIR, it is not organism-specific

Pathway inference

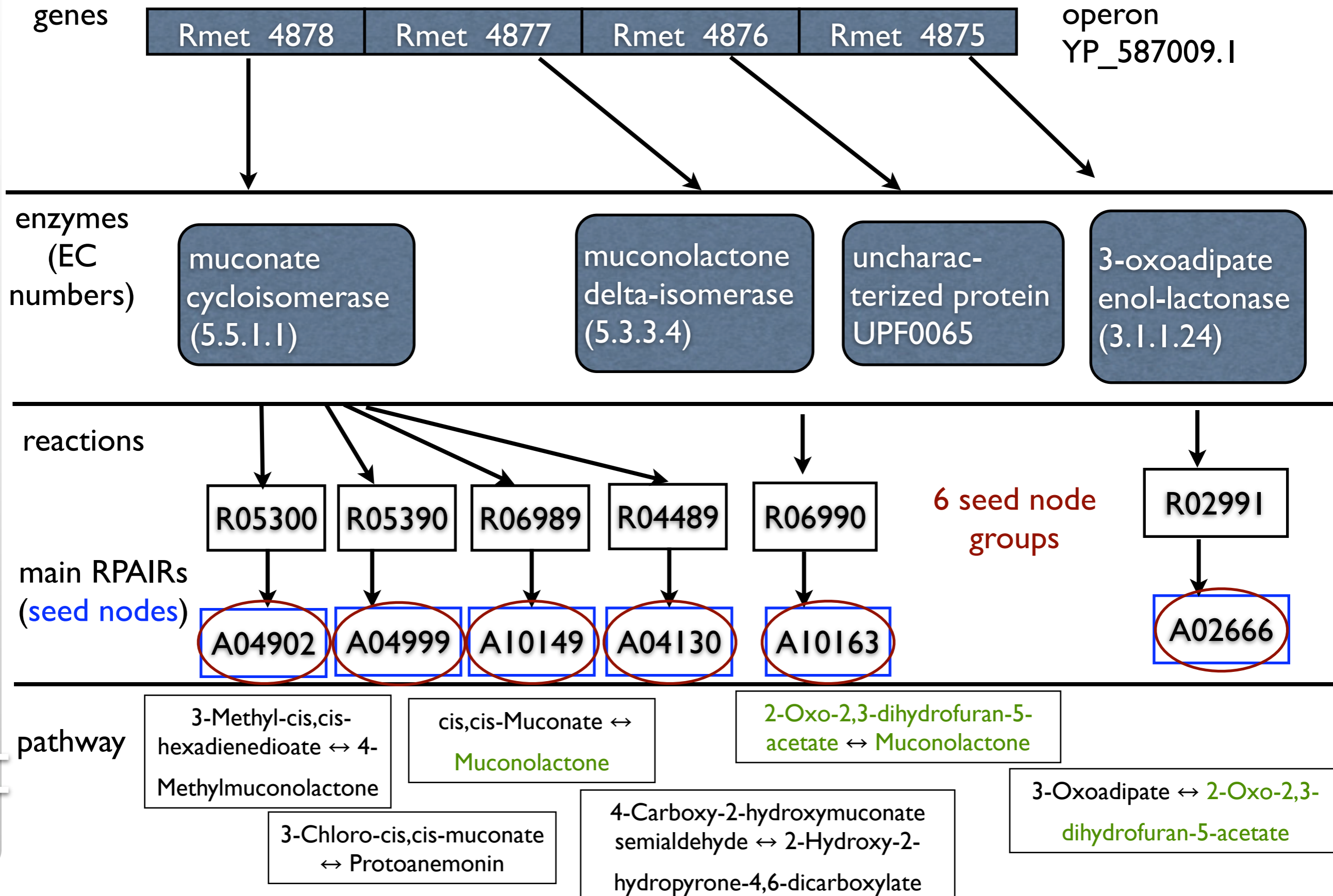
- hybrid algorithm
- 262 successful pathway inferences

1) R. Janky and J. van Helden: infer-operons (RSAT)

Analysis of *R. metallidurans* operons



Study case I: Obtaining reactions for genes



Application

Study case I: Pathway mapping

result of the KEGG pathway mapping tool

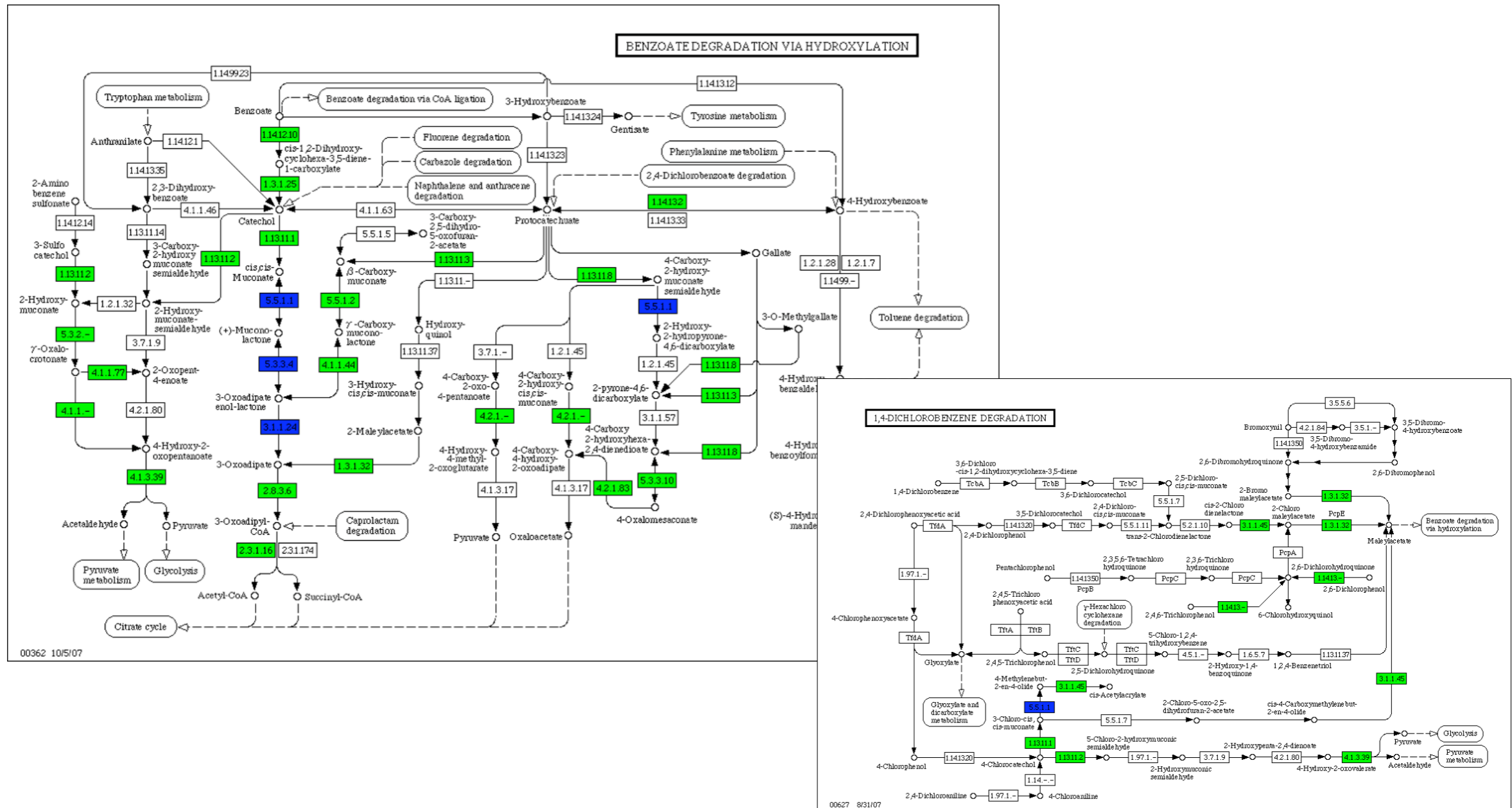
Rmet 4878

Rmet 4877

Rmet 4876

Rmet 4875

operon

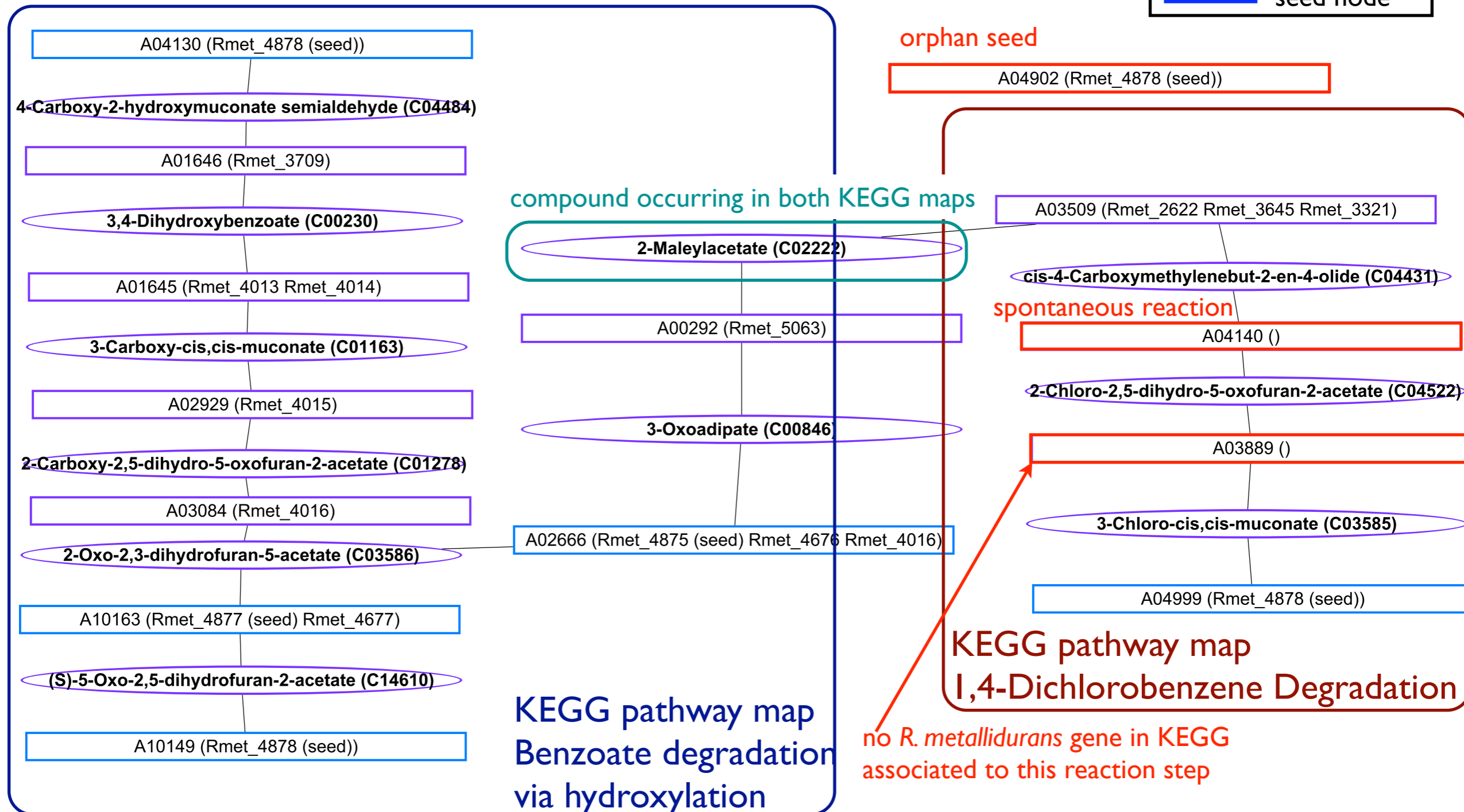
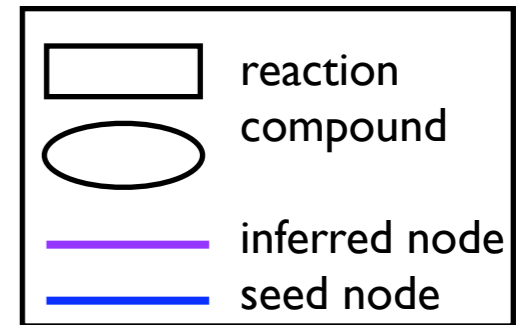


Application

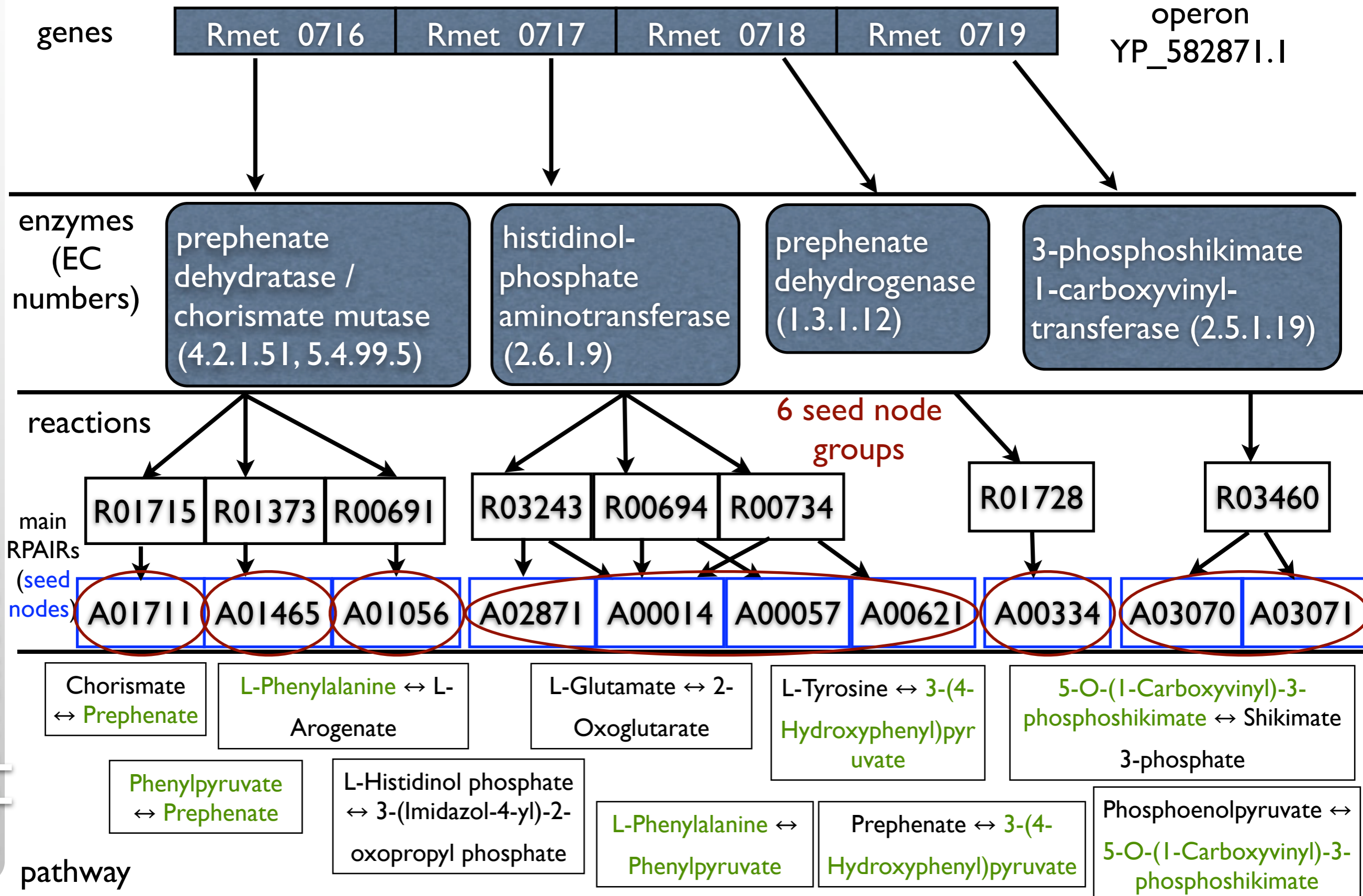
Study case I: Pathway inference

Inferred pathway combines 2 known pathways

Rmet 4878 Rmet 4877 Rmet 4876 Rmet 4875 operon



Study case II: Obtaining reactions for genes



Application

Study case II: Pathway mapping

result of the KEGG pathway mapping tool

Rmet 0716

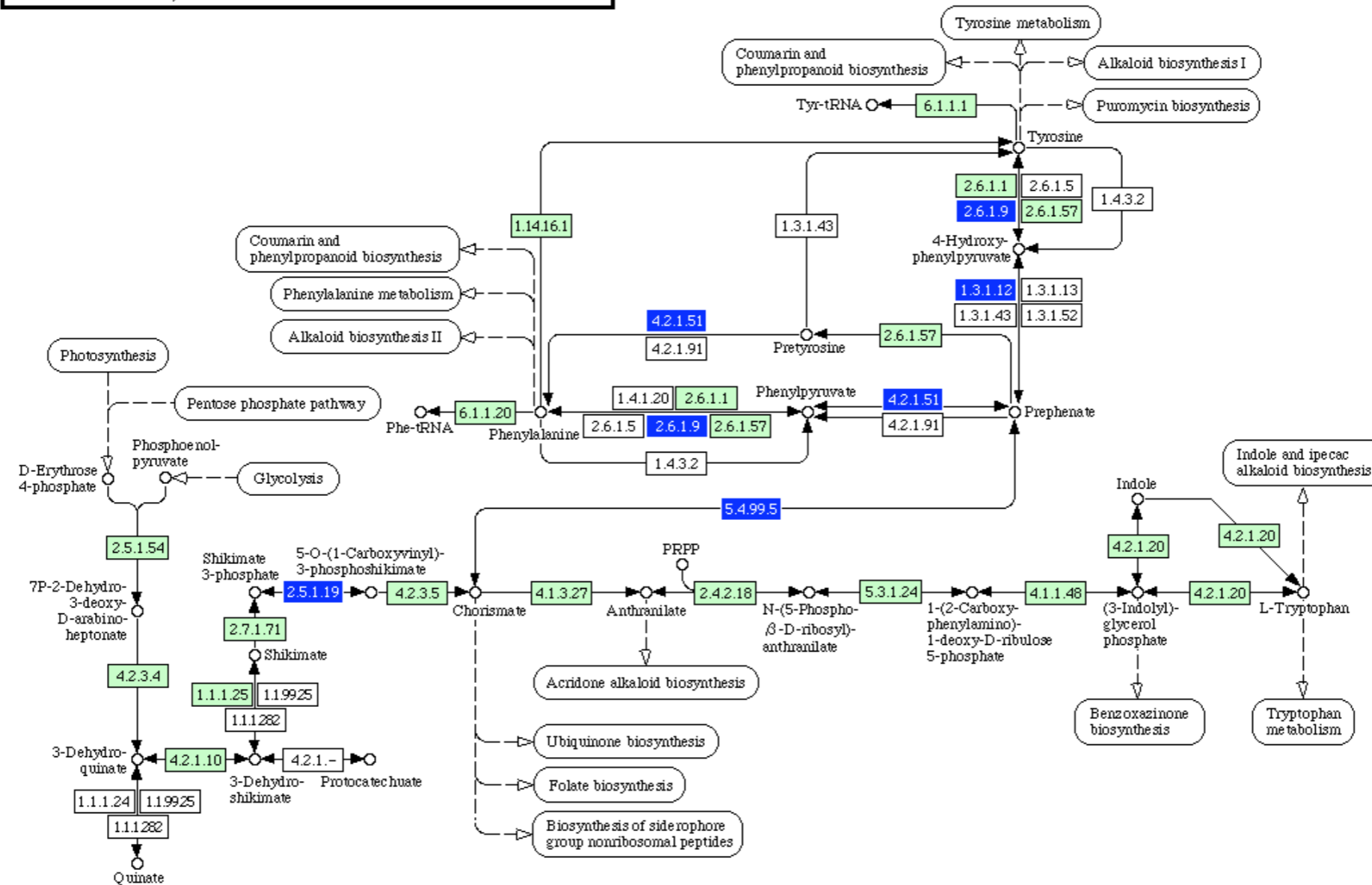
Rmet 0717

Rmet 0718

Rmet 0719

operon

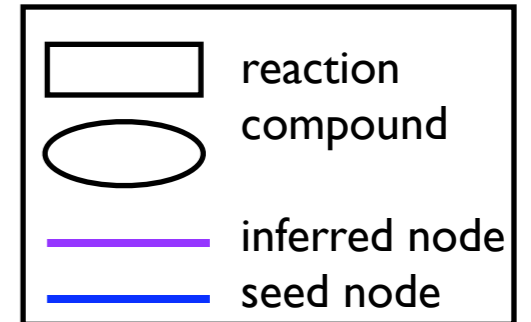
PHENYLALANINE, TYROSINE AND TRYPTOPHAN BIOSYNTHESIS



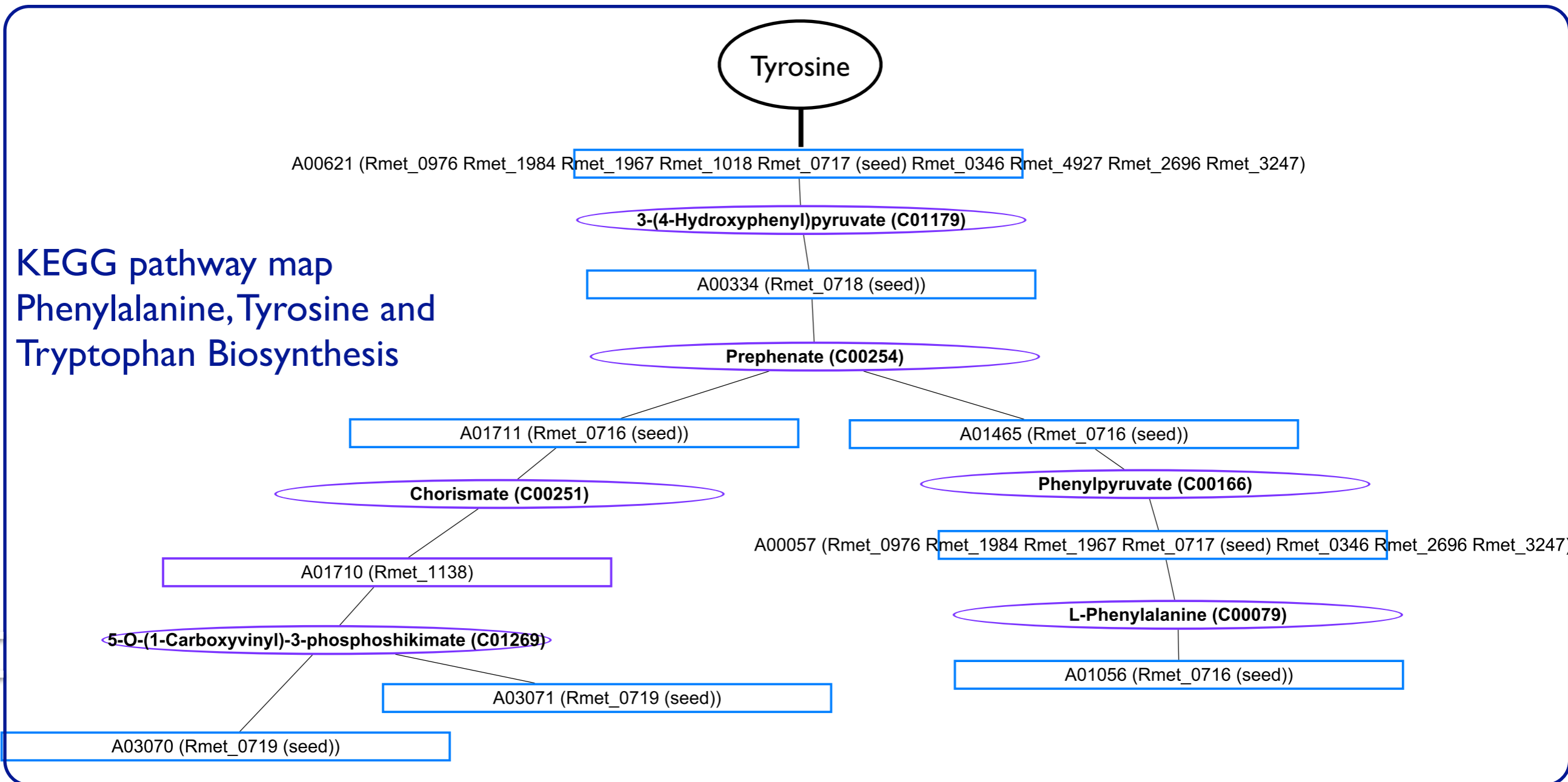
Study case II: Pathway inference

Inferred pathway corresponds to a known pathway

Rmet 0716 Rmet 0717 Rmet 0718 Rmet 0719 operon



KEGG pathway map
Phenylalanine, Tyrosine and
Tryptophan Biosynthesis



Limitations of pathway inference

- directions of reactions cannot be inferred (metabolic graph is undirected or includes both directions for each reaction)
- inferring densely interconnected regions of metabolism (e.g. glycolysis, TCA cycle) with high accuracy requires many seeds

Conclusion

- combination of kWalks and pair-wise k -shortest paths in the hybrid approach yields highest accuracies
- application to biological data set (operons of *R. metallidurans*): inference of relevant metabolic pathways that consist mostly of known pathways or their combination

Next steps

- test Steiner tree algorithms in combination with kWalks (work in progress)
- apply pathway inference to other biological data sets (micro-array data from *R. metallidurans* and *S. cerevisiae*)
- make pathway inference available as Web Service

Acknowledgements



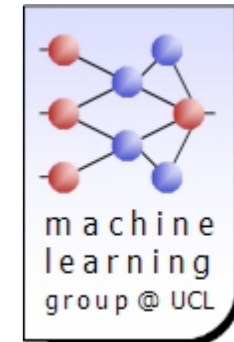
Jacques van Helden (supervisor)
Didier Croes
BiGRe team

IBMM

Bruno André
Patrice Godard



Fabian Couche
Christian Lemer
Hassan Anerhour
Frédéric Fays
Olivier Hubaut
Simon De Keyzer



Pierre Dupont
Jérôme Callut
Yves Deville
Pierre Schaus
Jean-Noël Monette

The PhD grant of Karoline Faust is funded by the Actions de Recherche Concertées de la Communauté Française de Belgique (ARC grant number 04/09-307). The INGI-BiGRe collaboration is funded by the Région Wallonne de Belgique (projects aMAZE and TransMaze).

Availability

Two-end path finding in weighted KEGG RPAIR graph (incorporating main/side compound annotation):

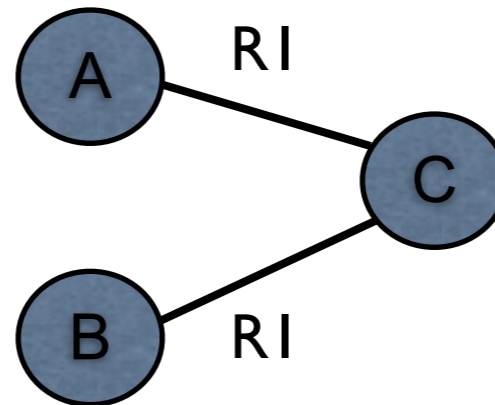
<http://rsat.ulb.ac.be/neat/> (Metabolic path finding)

Graph representation of metabolic data

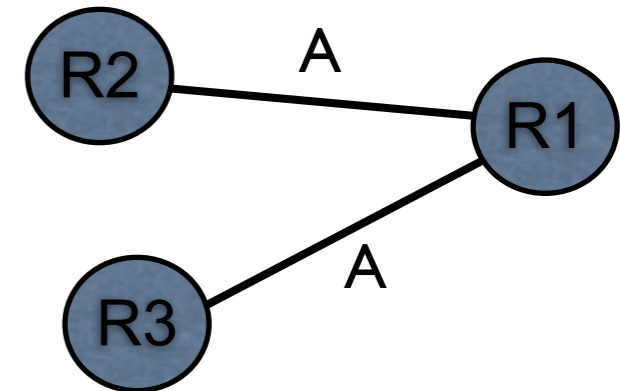
Why bipartite?

to avoid a compound or a reaction to be represented in the metabolic graph multiple times

graphs with only one node set:



reaction R1 is represented by several edges

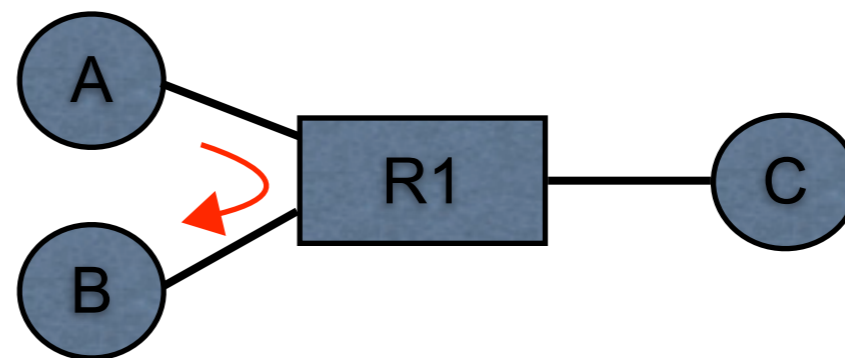


compound A is represented by several edges

Why directed?

to avoid paths going from educt to educt (or from product to product) of the same reaction

undirected graphs:



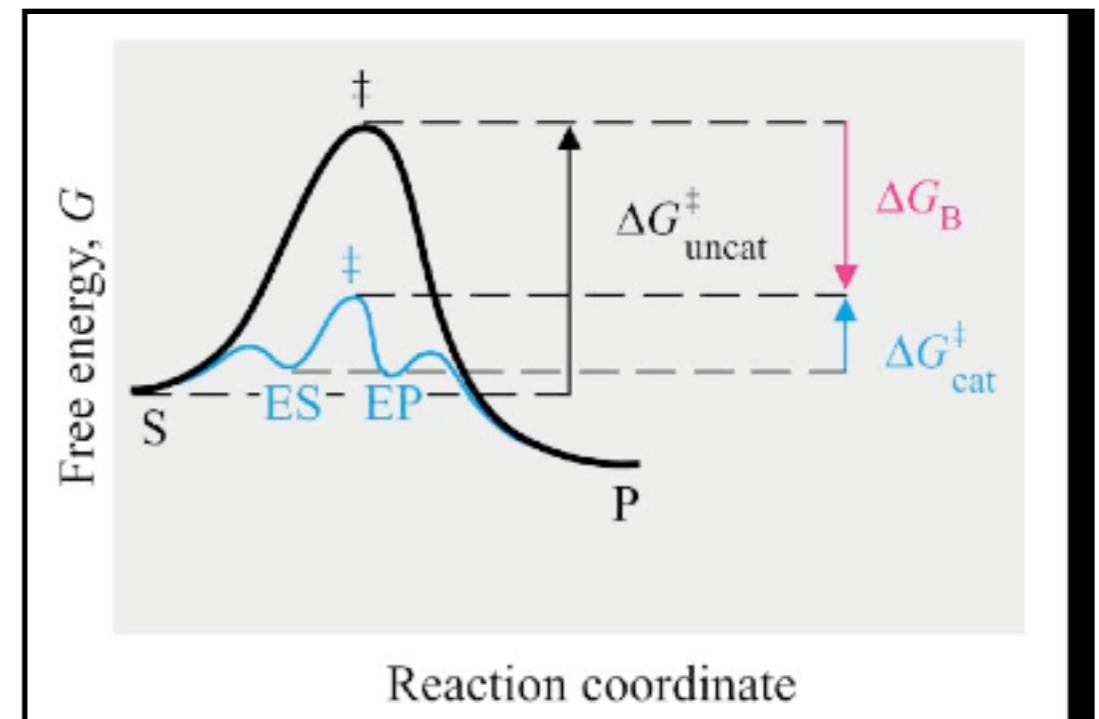
Why weighted?

to avoid highly connected compounds

Treatment of reaction directionality

- two ways to treat reaction directionality:
 - represent the reaction direction as annotated in the source database
 - consider that all the reactions can occur in both directions
- free energy ΔG depends on temperature T as well as on the product and substrate concentration ratio and the standard free energy ΔG°
- these parameters are known for only a few reactions - directed metabolic graph therefore contains direct and reverse direction for each reaction

enzymes don't alter the equilibrium of substrate and product concentrations, instead they speed up attainment of equilibria:



$$\Delta G = \Delta G^\circ + RT \ln\left(\frac{[\text{product}_1] \dots [\text{product}_m]}{[\text{educt}_1] \dots [\text{educt}_n]}\right)$$

image source: <http://www.biology.buffalo.edu/courses/bio401/KiongHo/Lecture32.pdf>

Weighting schemes

Node weighting schemes

compound node: degree or unit weight (1)

reaction node: unit weight (1)

Arc weight computation pair-wise k -shortest paths

- weight of arc a : mean of weight of head node n_h and weight of tail node n_t

$$w(a) = (w(n_h) + w(n_t)) / 2$$

Arc weight computation k Walks

- weight of arc a : inverse mean of weight of head node n_h and weight of tail node n_t :

$$w(a) = 2 / (w(n_h) + w(n_t))$$

Inflation of arc weight by inflation factor z :

$$w(a)^z$$

Construction of KEGG RPAIR graph I

- KEGG RPAIR: database of manually compiled reactant pairs that covers 6,261 reactions (1,128 reactions are not covered)
- reactant pairs: reaction-specific main/side compound annotation
- reactant pairs are classified as **main, cofac, trans, ligase or leave**

KEGG REACTION: R00256 Help

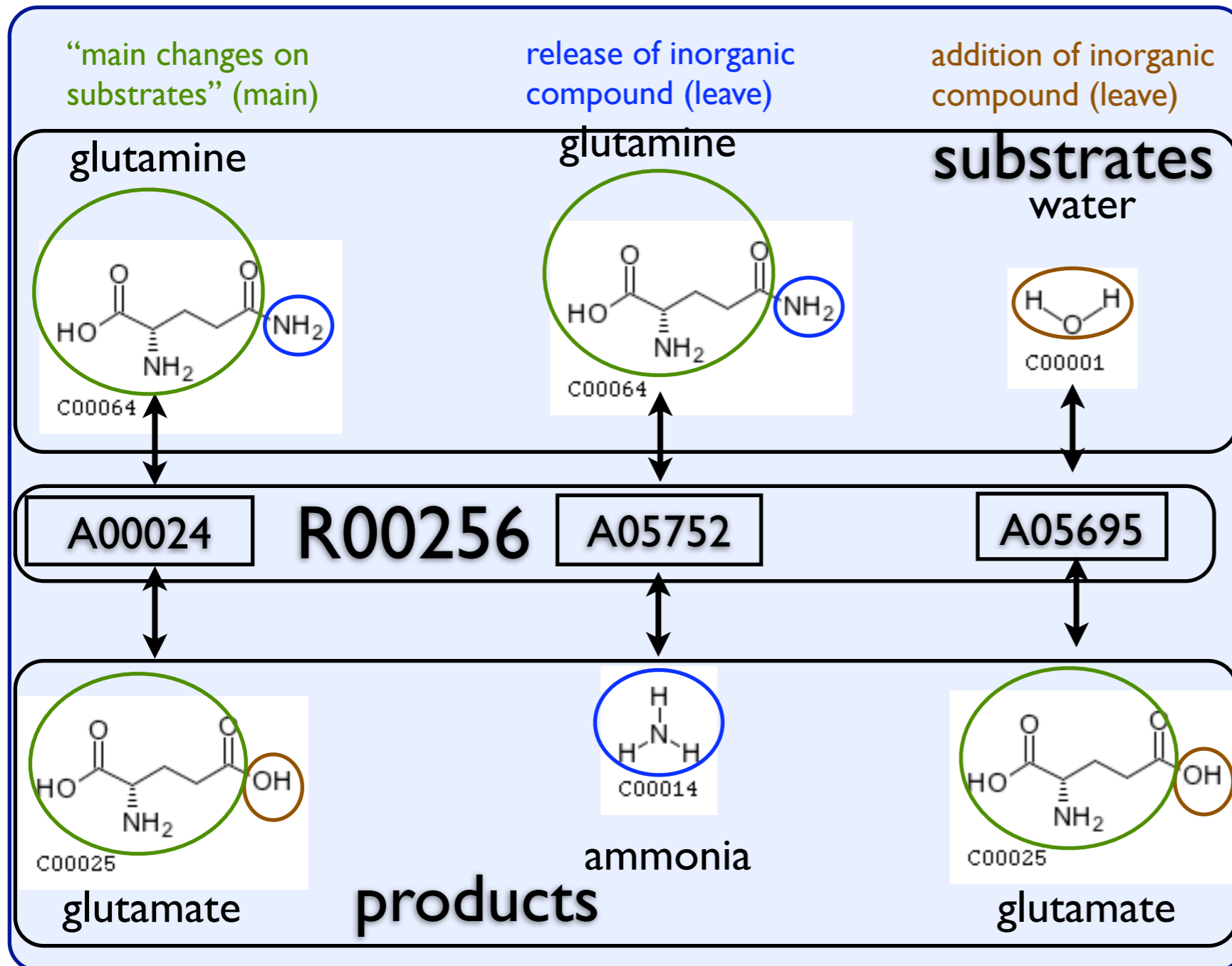
Entry	R00256	Reaction
Name	L-Glutamine amidohydrolase	
Definition	L-Glutamine + H ₂ O \rightleftharpoons L-Glutamate + NH ₃	
Equation	C00064 + C00001 \rightleftharpoons C00025 + C00014	
RPair	RP: A00024 C00025_C00064 main RP: A05695 C00001_C00025 leave RP: A05752 C00014_C00064 leave	

Kotera, M., Hattori, M., Oh, M.-A., Yamamoto, R., Komeno, T., Yabuzaki, J., Tonomura, K., Goto, S., and Kanehisa, M. (2004). "RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions" *Genome Informatics* 15. M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya (2002). "The KEGG databases at GenomeNet." *Nucleic Acids Research* 30(1): 42-46.

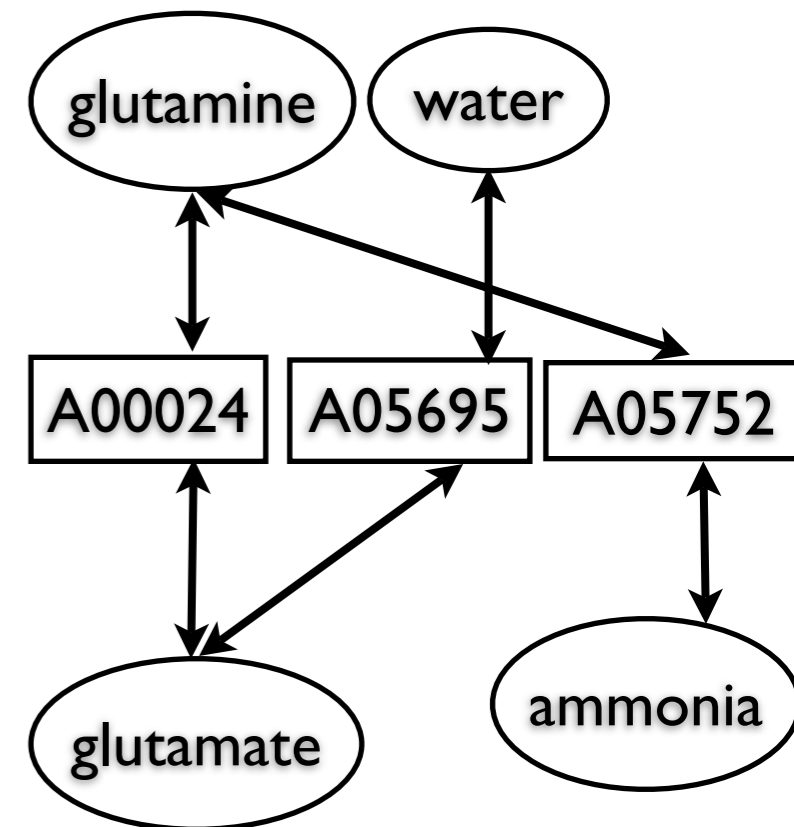
Construction of KEGG RPAIR graph II

graph constructed from all **reactant pairs** listed in KEGG and their associated **compounds**

reaction R00256 divided in its reactant pairs



presentation of reaction R00256 in the KEGG RPAIR graph



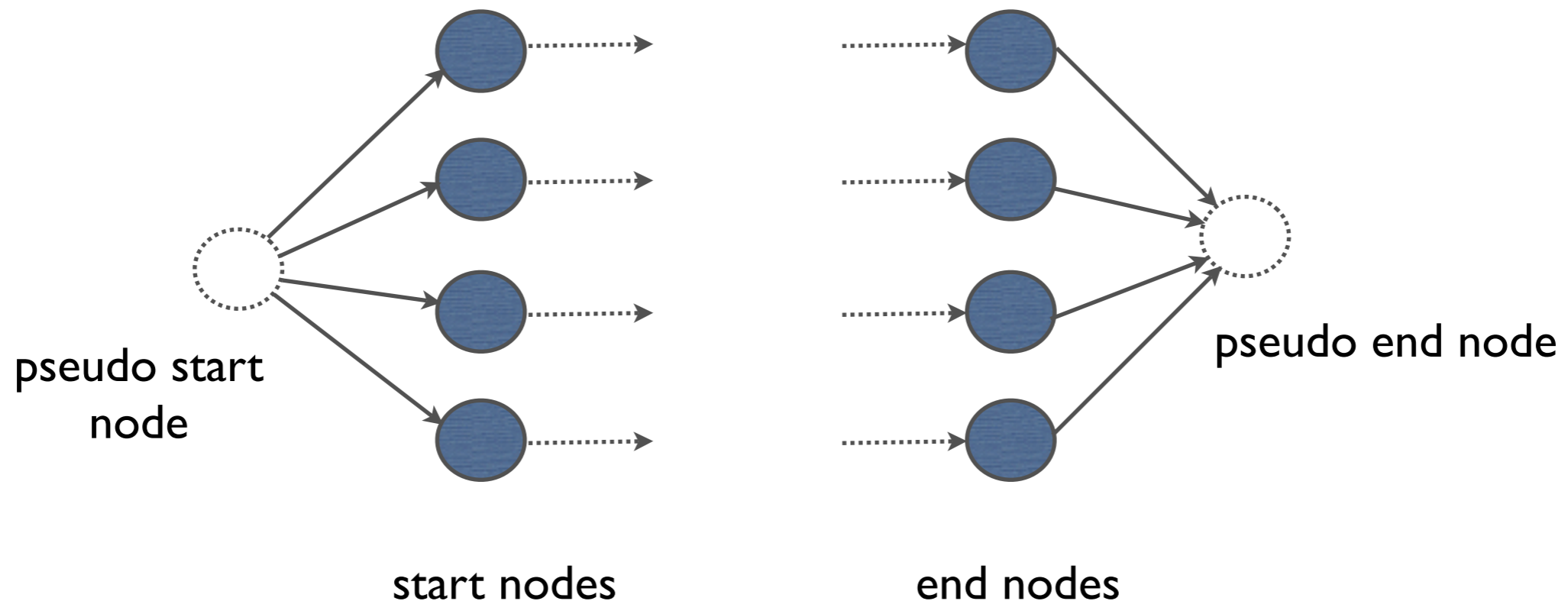
Treatment of seed node groups

kWalks

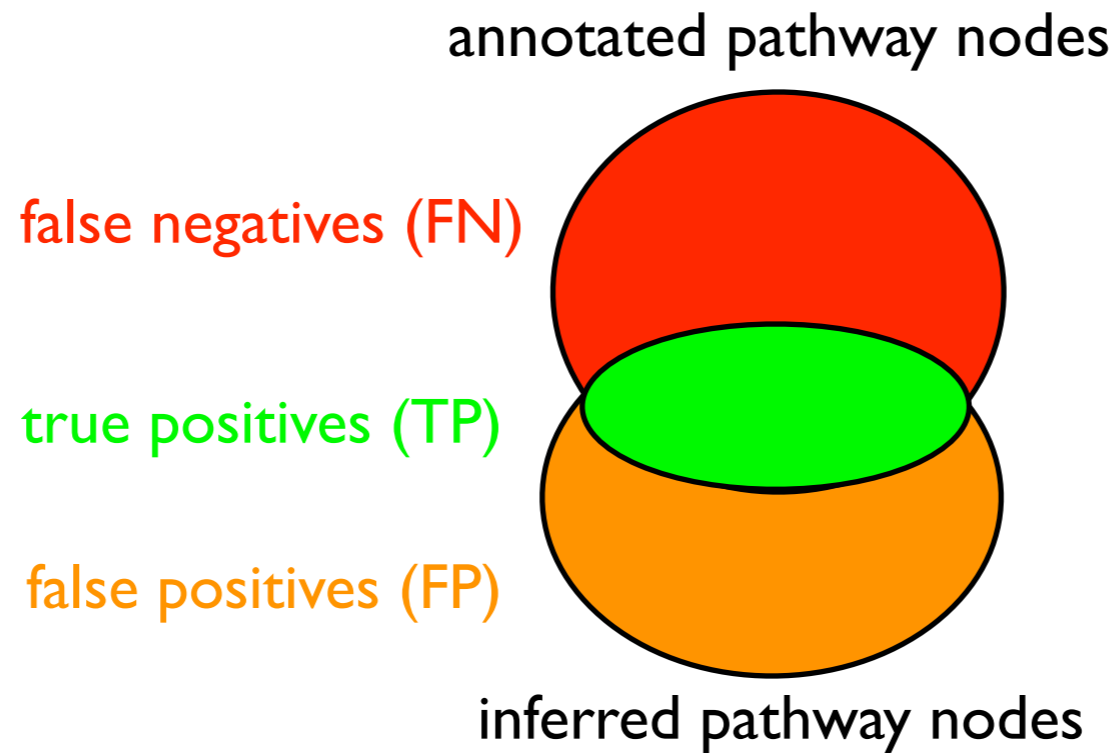
- random walks start in any node of group A and end in any node of group B

Pair-wise k -shortest paths

- multiple to multiple end path finding by introducing pseudo start and end nodes



Accuracy of pathway inference



sensitivity S_n : $TP / (TP + FN)$

positive predictive value PPV: $TP / (TP + FP)$

arithmetic accuracy: $(S_n + PPV) / 2$

geometric accuracy: $\sqrt{S \cdot PPV}$