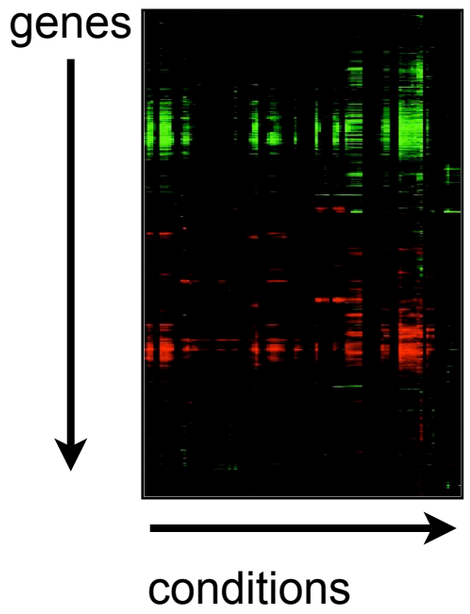
Inference of pathways from metabolic networks by subgraph extraction

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

*Laboratoire de Bioinformatique des Génomes et Réseaux (formerly SCMBB), ULB
http://www.scmbb.ulb.ac.be
+UCL Machine Learning Group
http://www.ucl.ac.be/mlg/

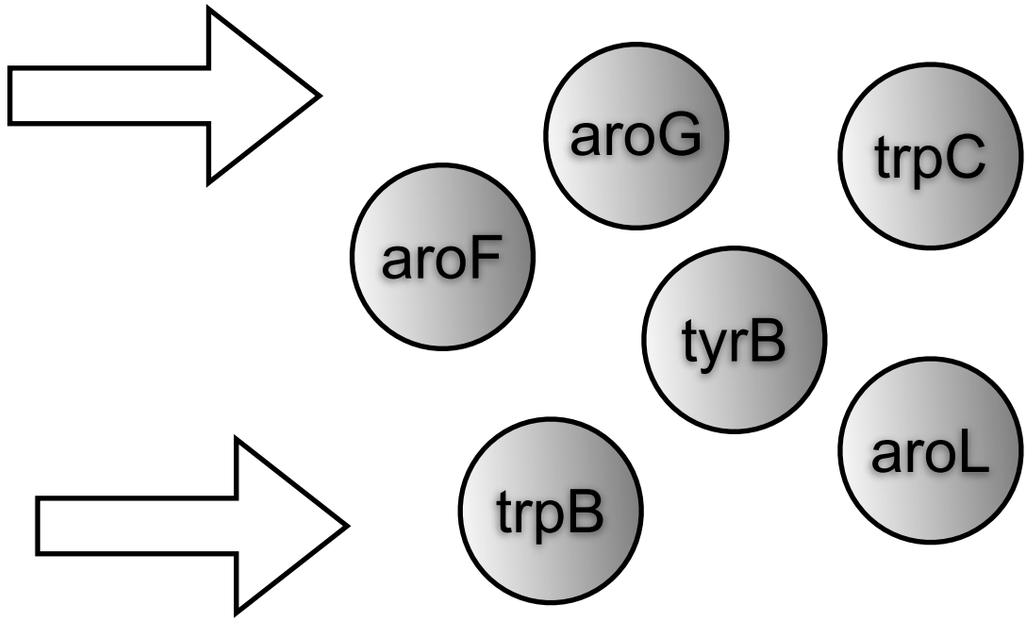
1. Motivation - Link genes assumed to be functionally related

microarray data

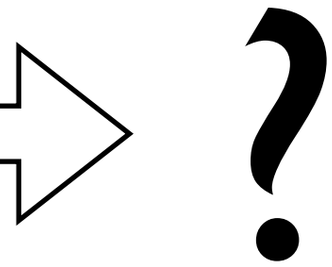


other data sets yielding gene groups assumed to be functionally related (same operon, co-regulation, ...)

enzyme-coding set of genes



pathway(s)



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

1. Motivation - Pathway mapping

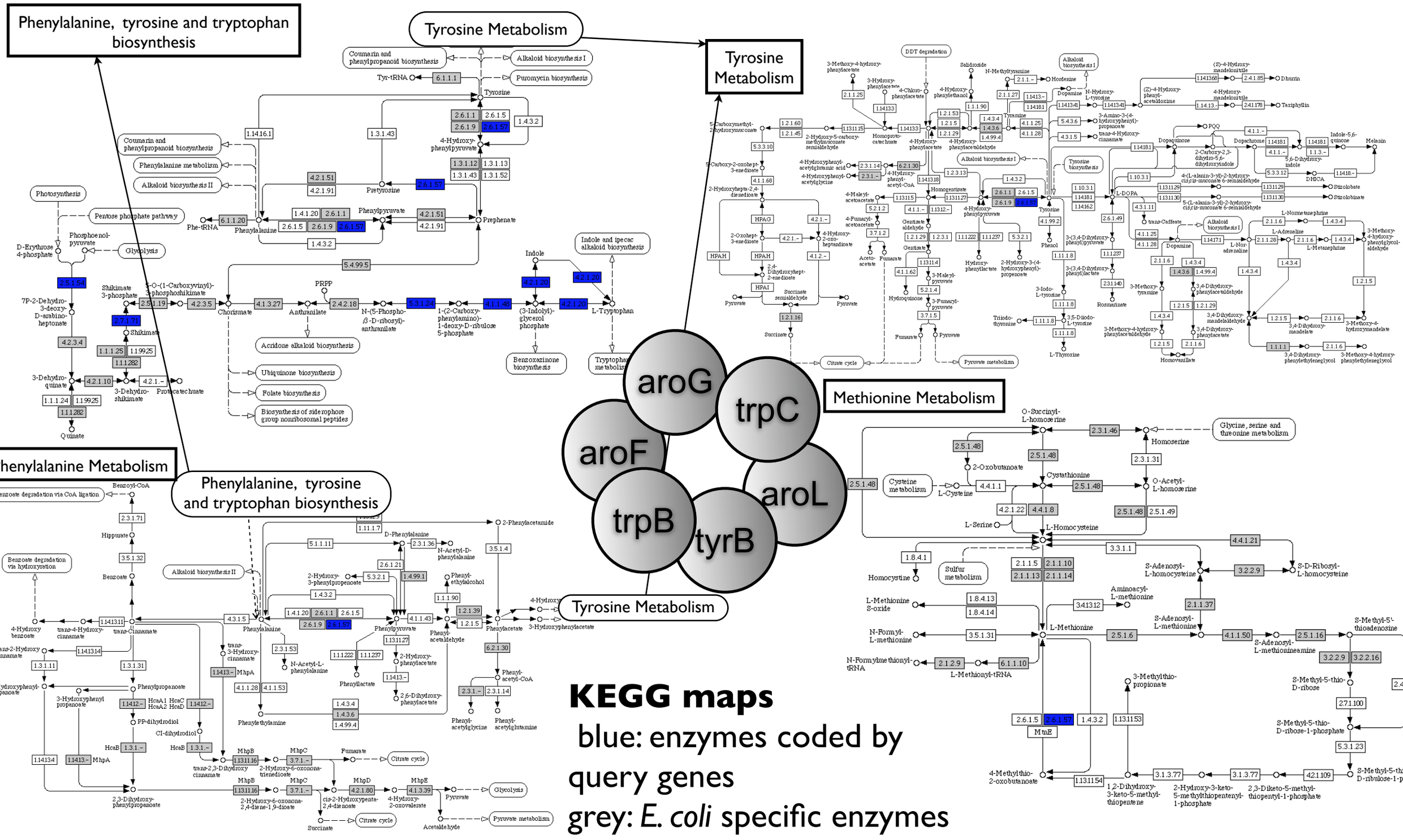
KEGG pathway mapping

Pathway Search Result

- **eco00400 Phenylalanine, tyrosine and tryptophan biosynthesis**
  b0388 aroL; shikimate kinase II [EC:2.7.1.71] [SP:AROL_ECOLI]
  b0754 aroG; 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase (DAHP synthetase, phenylalanine repressible) [EC:2.5.1.54] [SP:AROG_ECOLI]
  b1261 trpB; tryptophan synthase, beta protein [EC:4.2.1.20] [SP:TRPB_ECOLI]
  b1262 trpC; fused indole-3-glycerolphosphate synthetase/N-(5-phosphoribosyl)anthranilate isomerase [EC:5.3.1.24 4.1.1.48] [SP:TRPC_ECOLI]
  b2601 aroF; 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase (DAHP synthetase), tyrosine-repressible [EC:2.5.1.54] [SP:AROF_ECOLI]
  b4054 tyrB; tyrosine aminotransferase, tyrosine repressible [EC:2.6.1.57] [SP:TYRB_ECOLI]
- **eco02020 Two-component system - General**
  b1261 trpB; tryptophan synthase, beta protein [EC:4.2.1.20] [SP:TRPB_ECOLI]
  b1262 trpC; fused indole-3-glycerolphosphate synthetase/N-(5-phosphoribosyl)anthranilate isomerase [EC:5.3.1.24 4.1.1.48] [SP:TRPC_ECOLI]
- **eco00271 Methionine metabolism**
  b4054 tyrB; tyrosine aminotransferase, tyrosine repressible [EC:2.6.1.57] [SP:TYRB_ECOLI]
- **eco00350 Tyrosine metabolism**
  b4054 tyrB; tyrosine aminotransferase, tyrosine repressible [EC:2.6.1.57] [SP:TYRB_ECOLI]
- **eco00360 Phenylalanine metabolism**
  b4054 tyrB; tyrosine aminotransferase, tyrosine repressible [EC:2.6.1.57] [SP:TYRB_ECOLI]
- **eco00401 Novobiocin biosynthesis**
  b4054 tyrB; tyrosine aminotransferase, tyrosine repressible [EC:2.6.1.57] [SP:TYRB_ECOLI]
- **eco00950 Alkaloid biosynthesis I**
  b4054 tyrB; tyrosine aminotransferase, tyrosine repressible [EC:2.6.1.57] [SP:TYRB_ECOLI]

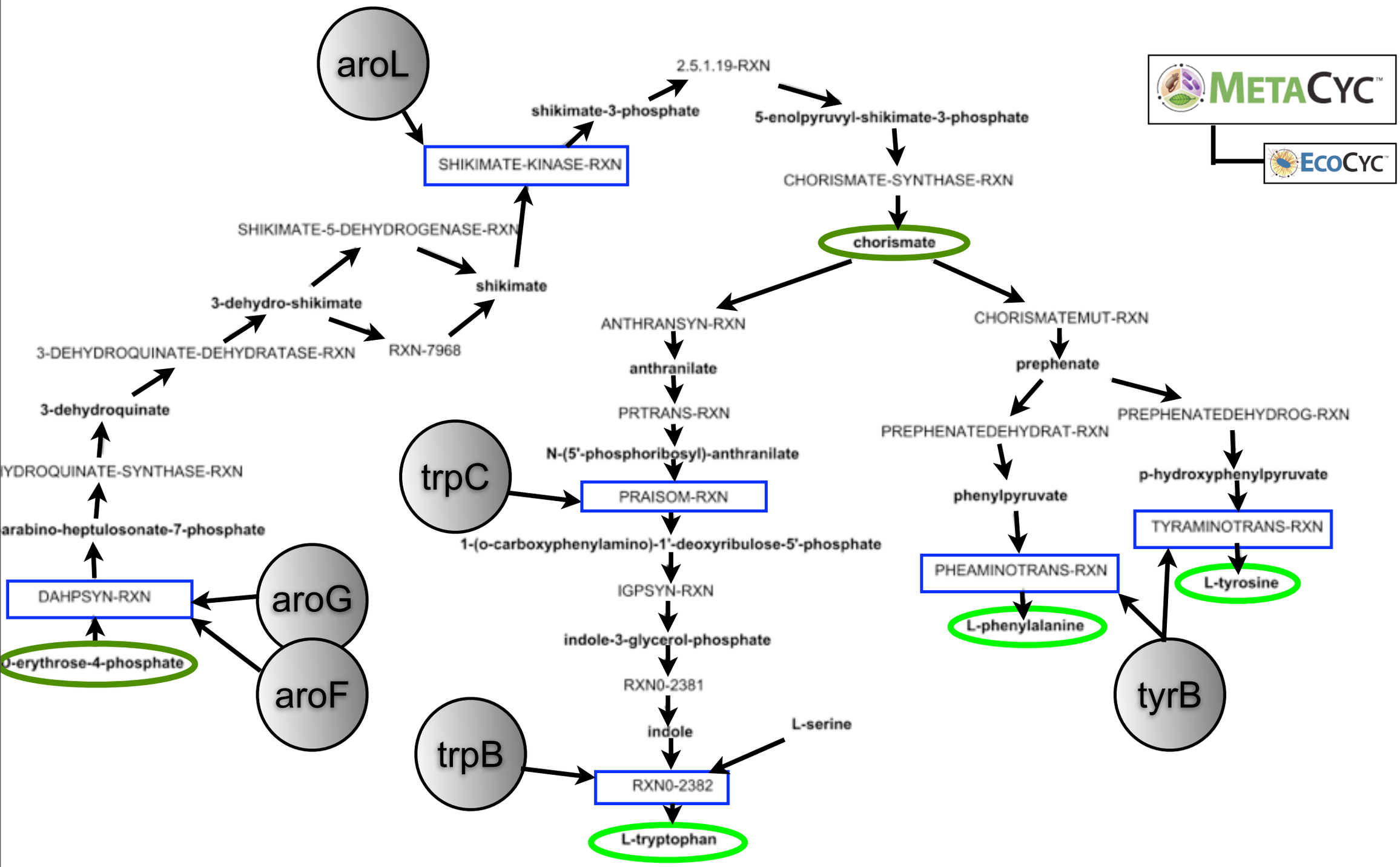
M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya. (2002). "The KEGG databases at GenomeNet", Nucleic Acids Research 30: 42-46.

1. Motivation - Result of pathway mapping



1. Motivation - Enzymes involved in known example pathway

superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis



C. Krieger et al. (2004). “MetaCyc: a multiorganism database of metabolic pathways and enzymes.” Nucleic Acids Research 32: D438-D442.

1. Motivation - Pathway mapping limitations

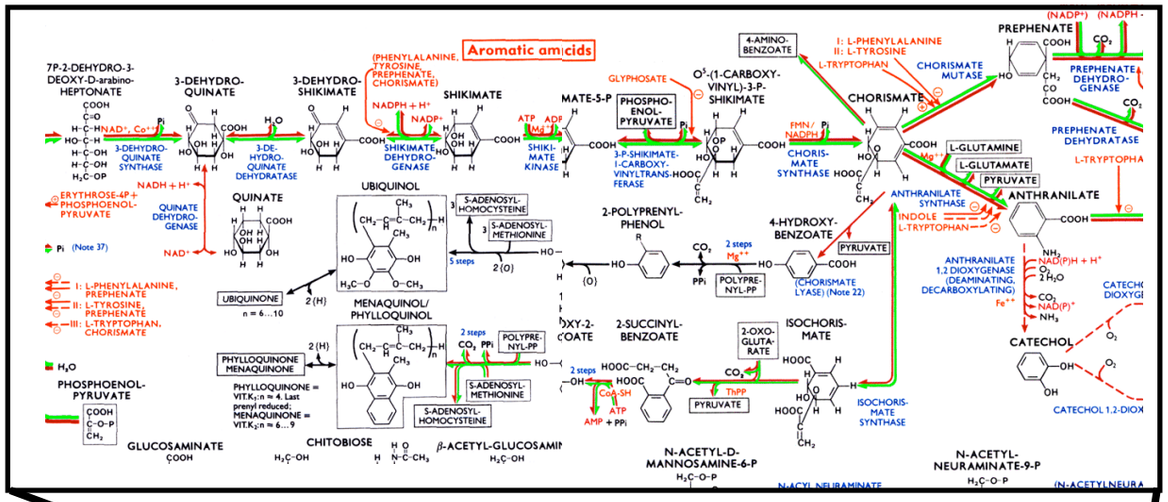
Why is pathway mapping not sufficient?

- pre-defined pathway set may be incomplete

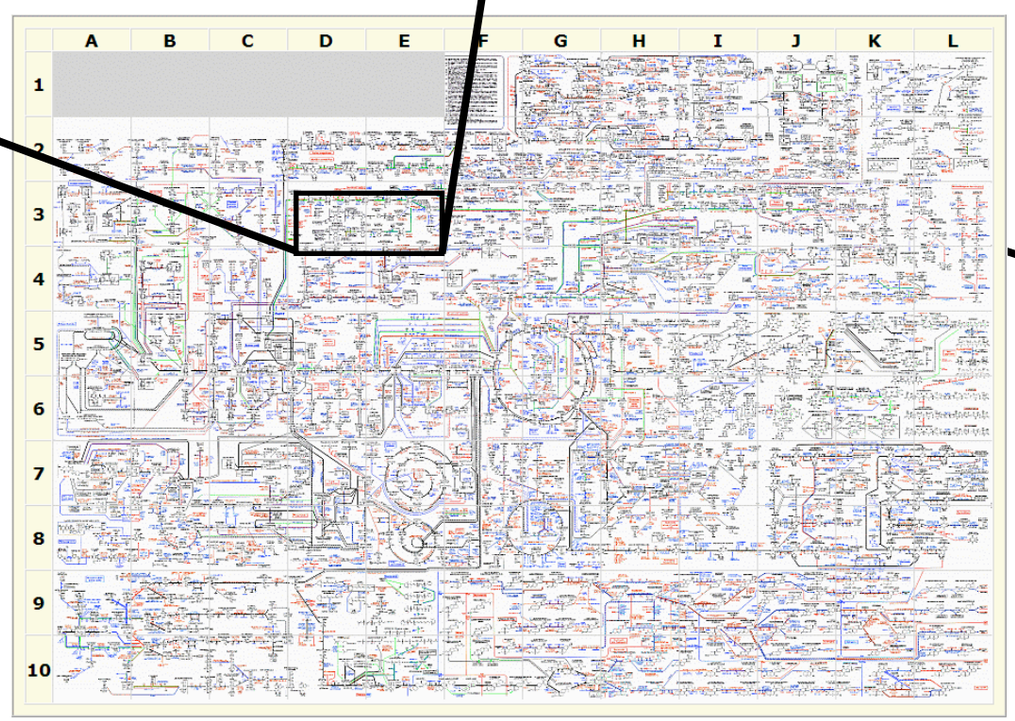
- mapping does not deal well with genes that map to several pre-defined pathways

- mapping does not allow variations or combinations of pathways

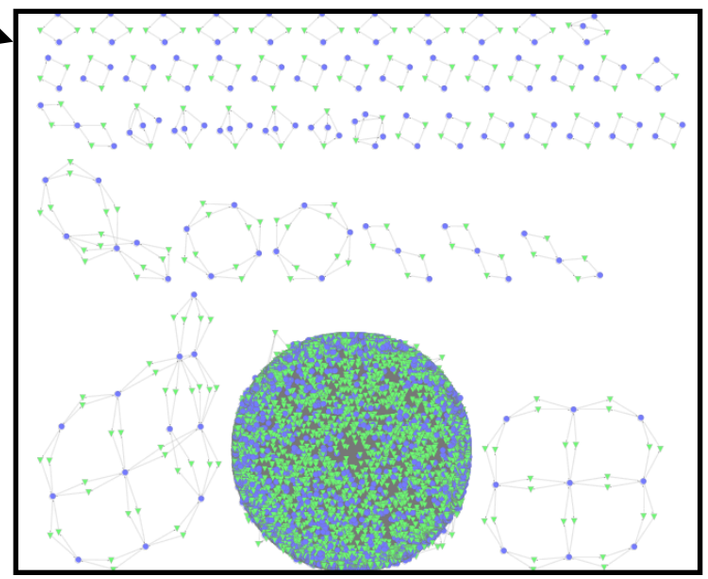
1. Motivation - Metabolic networks



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



biochemical pathways wall chart (Roche)



metabolic graph constructed from MetaCyc

2. Aim - Metabolic pathway inference

given a set of enzyme-coding genes, find meaningful metabolic pathways connecting them

J. van Helden, D. Gilbert, L. Wernisch, M. Schroeder, S. Wodak (2001) “Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data.” Lecture Notes in Computer Science, Vol. 2066, 147-165

2. Aim - Metabolic pathway inference

given a set of enzyme-coding genes, find meaningful metabolic pathways connecting them

functionally related genes

↓

enzyme-coding genes

↓

reactions associated to enzymes as seed nodes

J. van Helden, D. Gilbert, L. Wernisch, M. Schroeder, S. Wodak (2001) “Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data.” Lecture Notes in Computer Science, Vol. 2066, 147-165

2. Aim - Metabolic pathway inference

given a set of enzyme-coding genes, find meaningful metabolic pathways connecting them

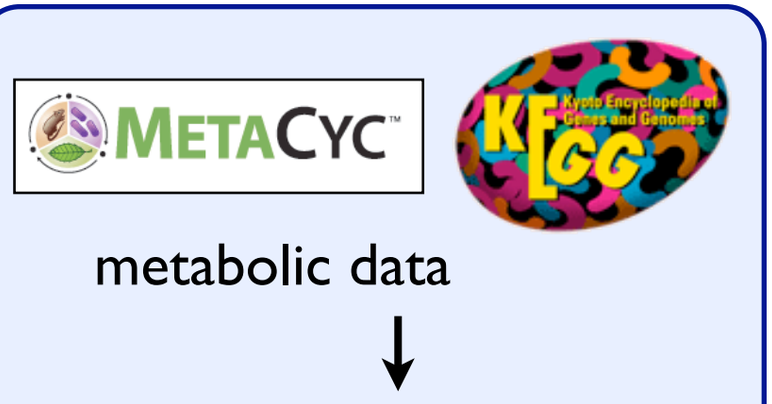
functionally related genes

↓

enzyme-coding genes

↓

reactions associated to enzymes as seed nodes



J. van Helden, D. Gilbert, L. Wernisch, M. Schroeder, S. Wodak (2001) "Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data." Lecture Notes in Computer Science, Vol. 2066, 147-165

2. Aim - Metabolic pathway inference

given a set of enzyme-coding genes, find meaningful metabolic pathways connecting them

functionally related genes

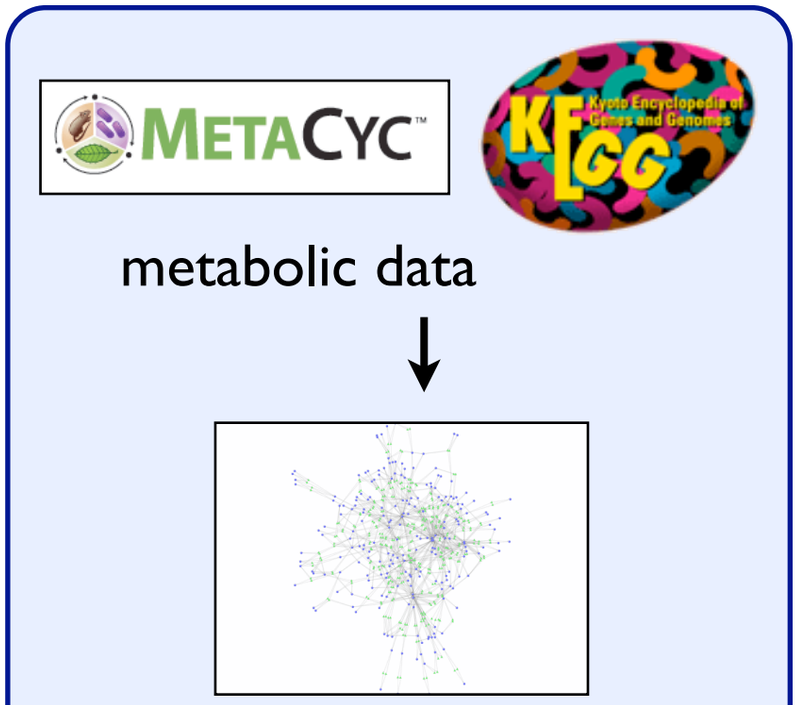
↓

enzyme-coding genes

↓

reactions associated to enzymes as seed nodes

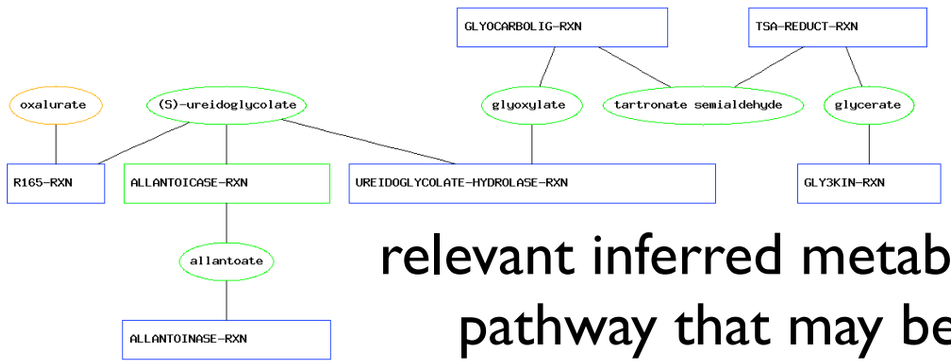
metabolic pathway inference



metabolic data

↓

metabolic graph (reaction and compound nodes)



relevant inferred metabolic pathway that may be validated experimentally

J. van Helden, D. Gilbert, L. Wernisch, M. Schroeder, S. Wodak (2001) “Application of Regulatory Sequence Analysis and Metabolic Network Analysis to the Interpretation of Gene Expression Data.” Lecture Notes in Computer Science, Vol. 2066, 147-165

3. Methods - Two-end metabolic path finding

Approach

- infer pathway given two seed nodes only using path finding (k shortest paths) algorithm

- problem: hub nodes (highly connected compounds such as ATP, H_2O etc.) favor biochemically irrelevant pathways

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.

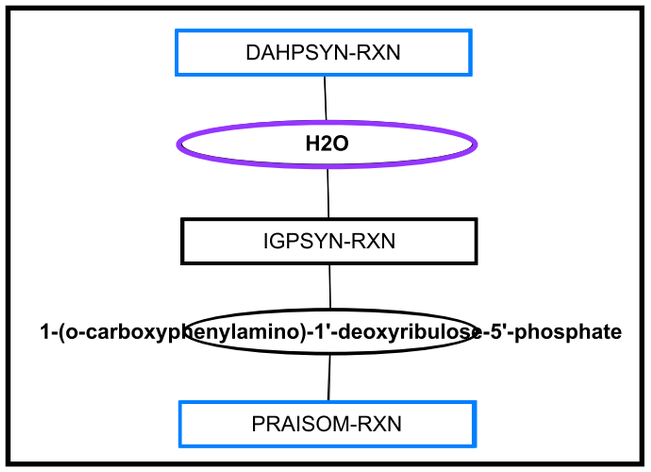
3. Methods - Two-end metabolic path finding

Approach

- infer pathway given two seed nodes only using path finding (k shortest paths) algorithm

- problem: hub nodes (highly connected compounds such as ATP, H_2O etc.) favor biochemically irrelevant pathways

unweighted graph (pathway traverses highly connected compound)



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.

3. Methods - Two-end metabolic path finding

Approach

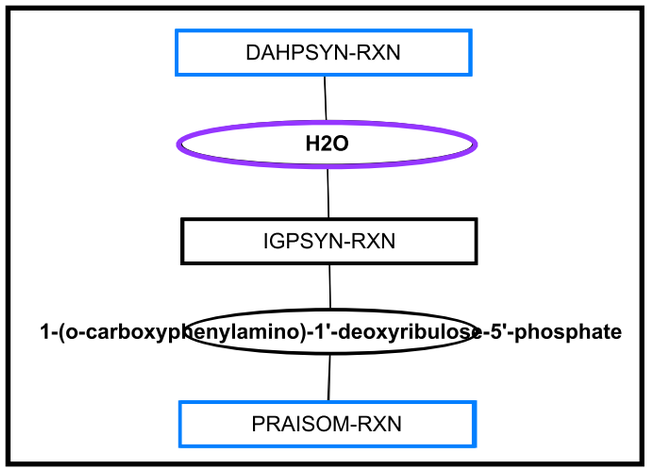
- infer pathway given two seed nodes only using path finding (k shortest paths) algorithm

- problem: hub nodes (highly connected compounds such as ATP, H_2O etc.) favor biochemically irrelevant pathways

- solution: weighted graph penalizing hubs

- weighted graph gives better results than either unweighted or filtered graph (hubs removed)

unweighted graph (pathway traverses highly connected compound)



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.

3. Methods - Two-end metabolic path finding

Approach

- infer pathway given two seed nodes only using path finding (k shortest paths) algorithm

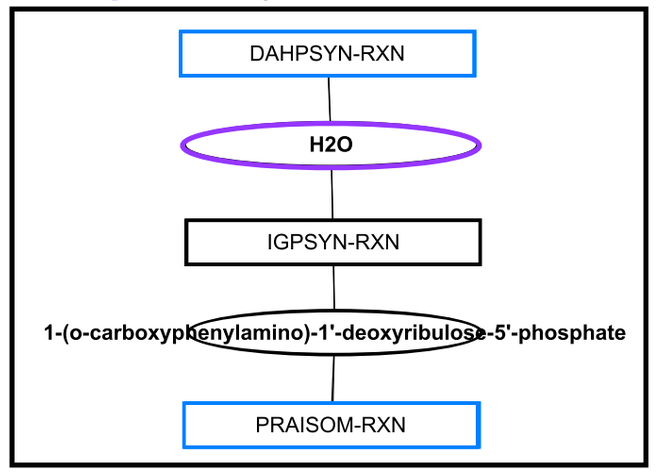
- problem: hub nodes (highly connected compounds such as ATP, H_2O etc.) favor biochemically irrelevant pathways

- solution: weighted graph penalizing hubs

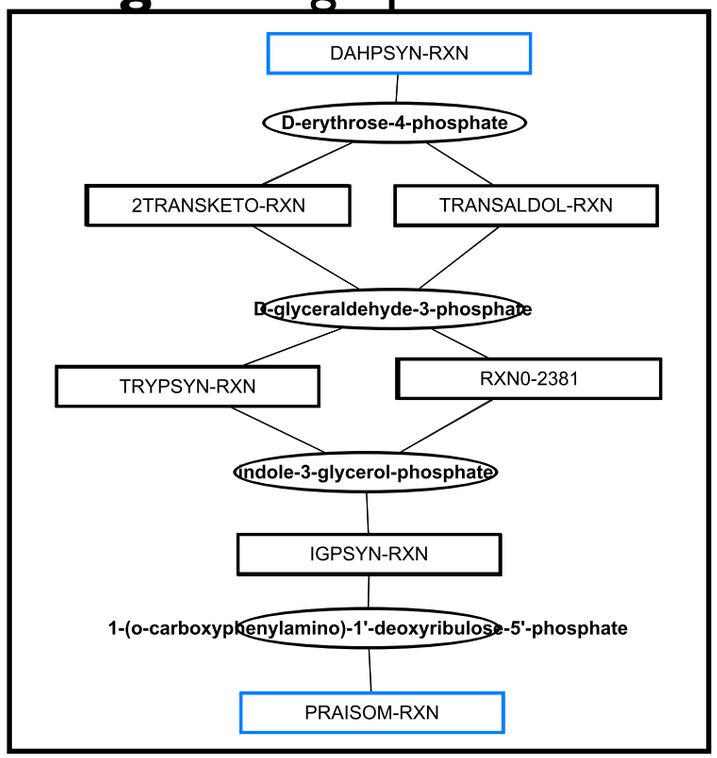
- weighted graph gives better results than either unweighted or filtered graph (hubs removed)

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.

unweighted graph (pathway traverses highly connected compound)



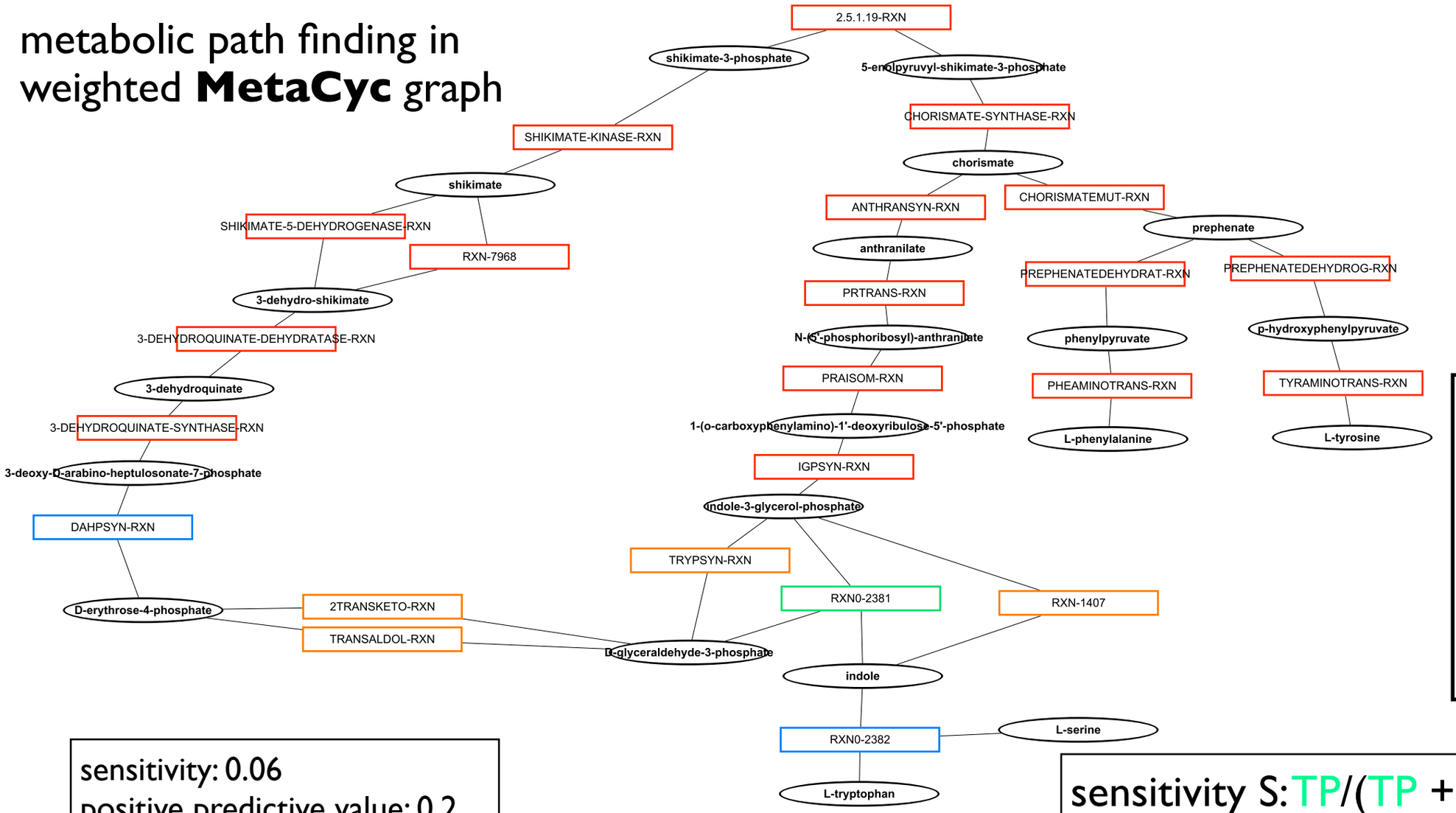
weighted graph



3. Methods - Definition of accuracy

reference: superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis

metabolic path finding in weighted **MetaCyc** graph



sensitivity: 0.06
positive predictive value: 0.2
arithmetic accuracy: 0.13
geometric accuracy: 0.11

seed reactions do not count as true positives

sensitivity S: TP/(TP + FN)

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

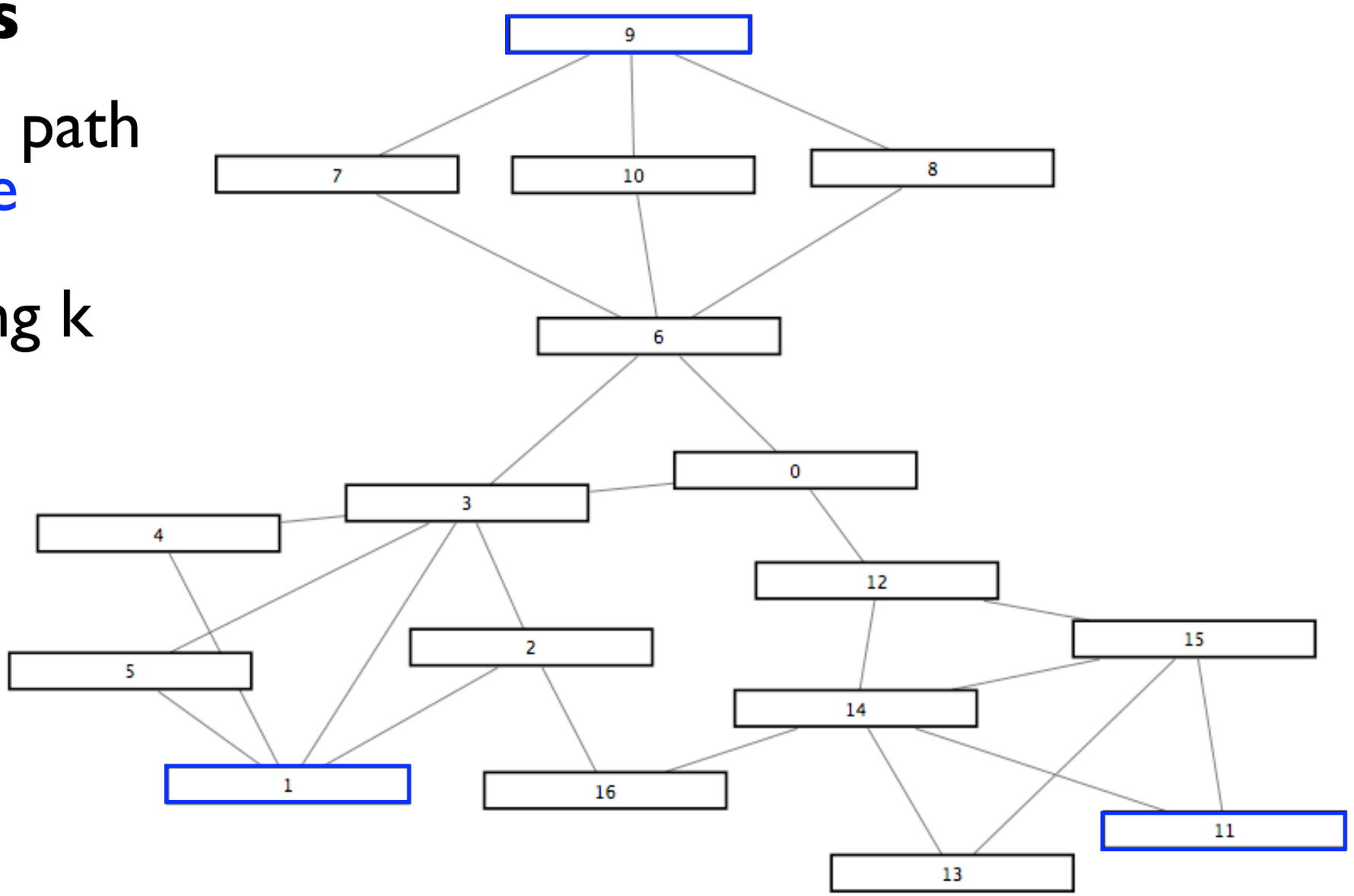
arithmetic accuracy: (S+PPV)/2

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

3. Methods - Multiple-end metabolic pathway inference

Pairwise k shortest paths

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

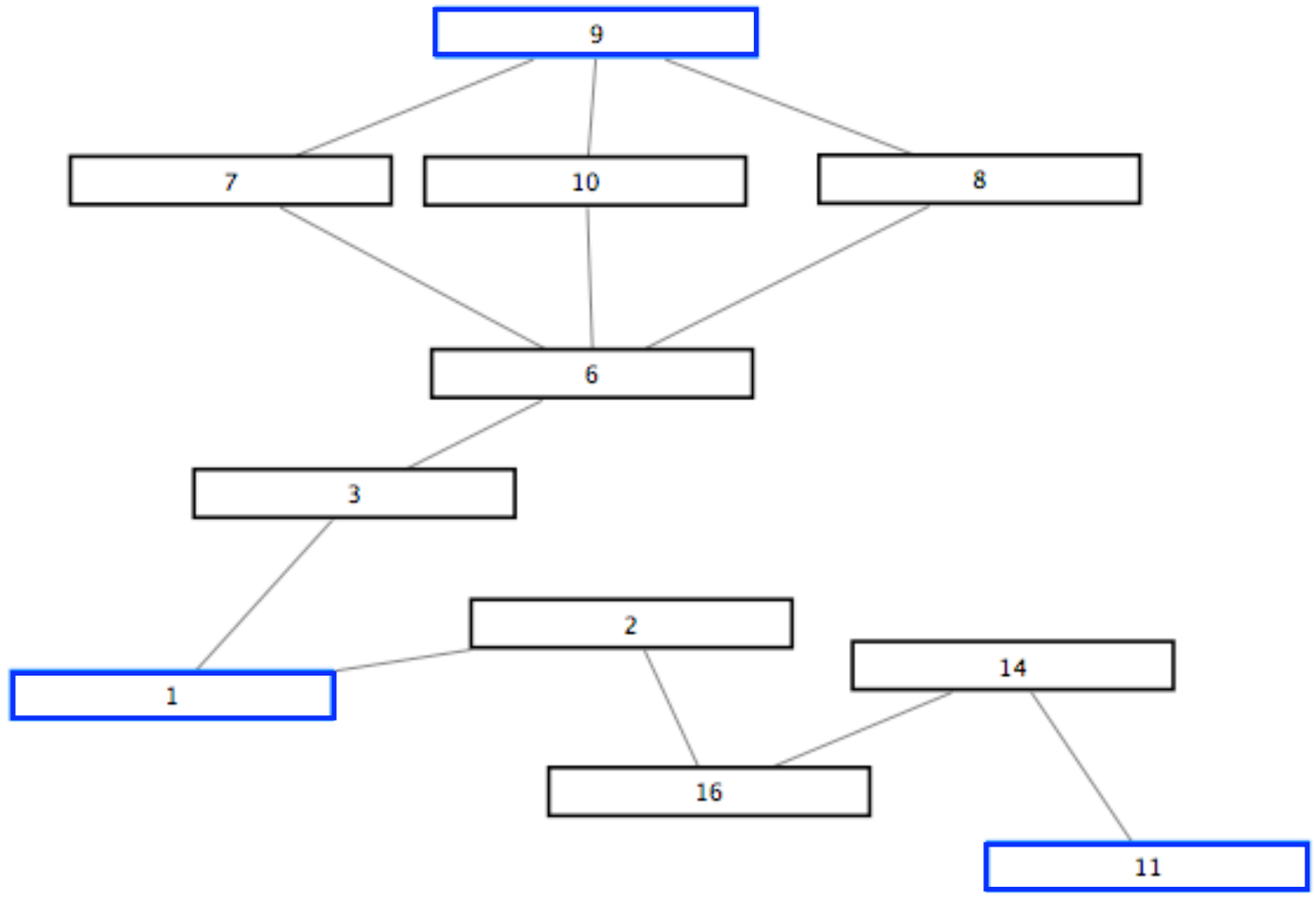


V.M. Jimenez and A. Marzal (1999). “Computing the K Shortest Paths: a New Algorithm and an Experimental Comparison.” Proc. 3rd Int. Worksh. Algorithm Engineering, Springer Verlag

3. Methods - Multiple-end metabolic pathway inference

Pairwise k shortest paths

- extract subgraph: unify lightest paths (of first rank) in the order of their weight until all seed nodes are connected

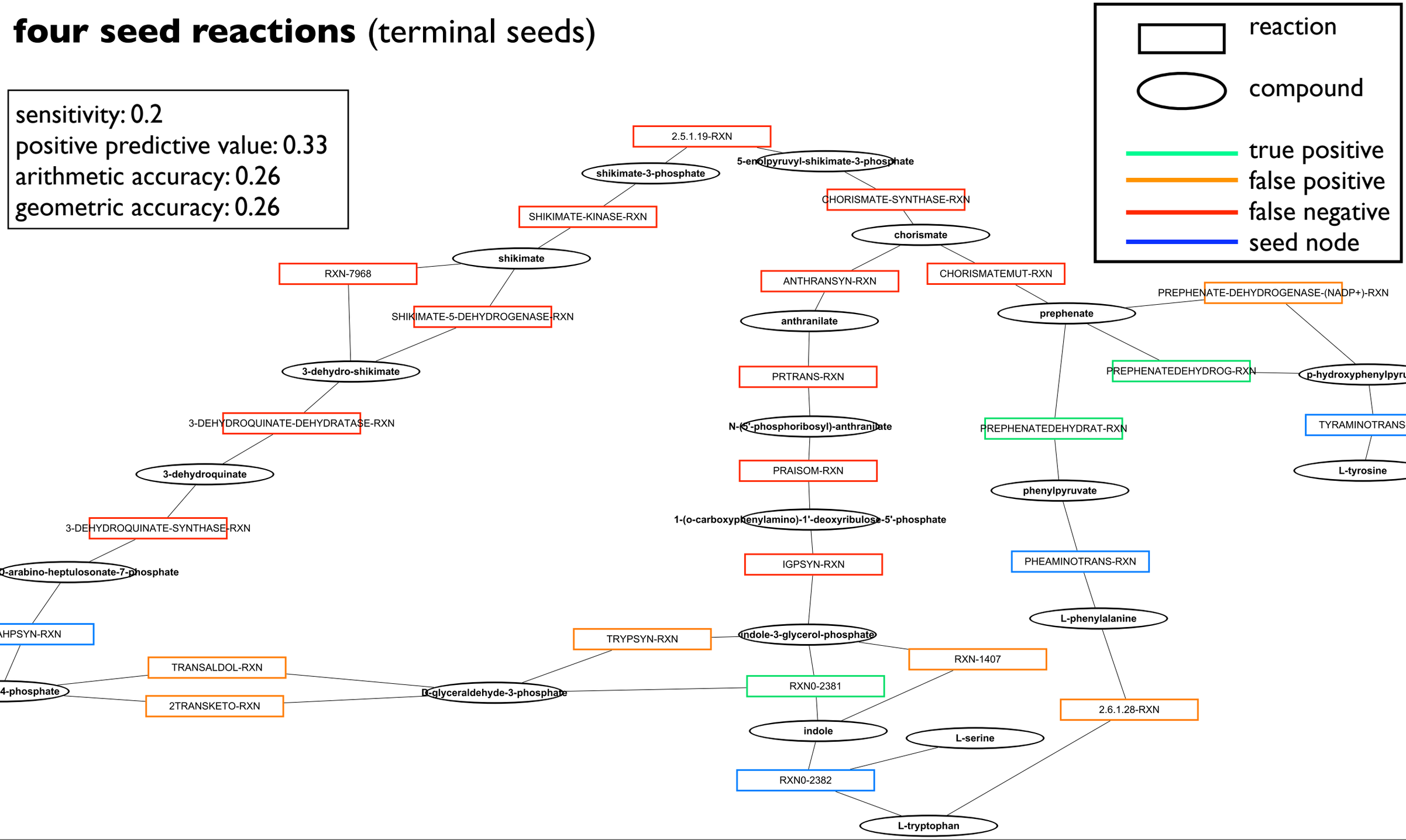


3. Methods - Pairwise k shortest paths in weighted MetaCyc graph

reference: superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis

four seed reactions (terminal seeds)

sensitivity: 0.2
positive predictive value: 0.33
arithmetic accuracy: 0.26
geometric accuracy: 0.26

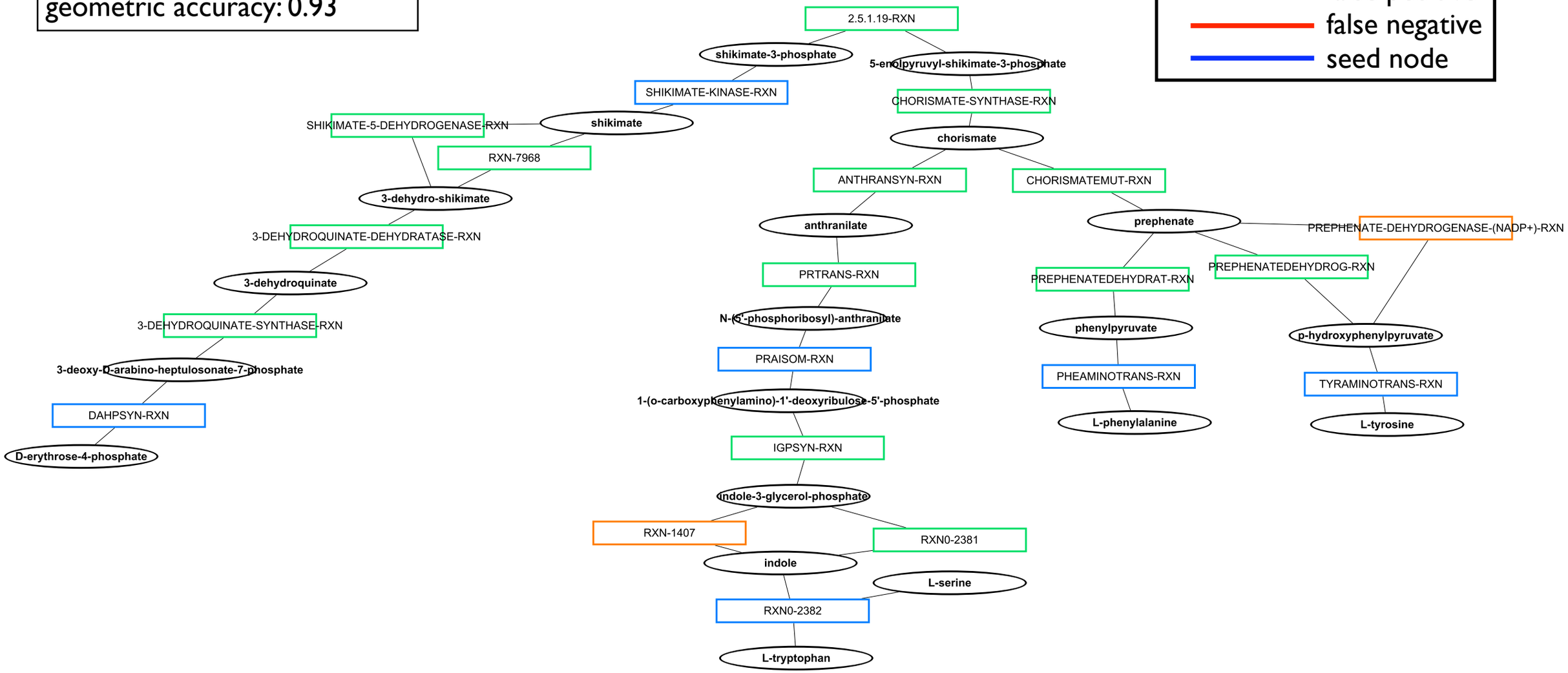


3. Methods - Pairwise k shortest paths in weighted MetaCyc graph

reference: superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis

six seed reactions

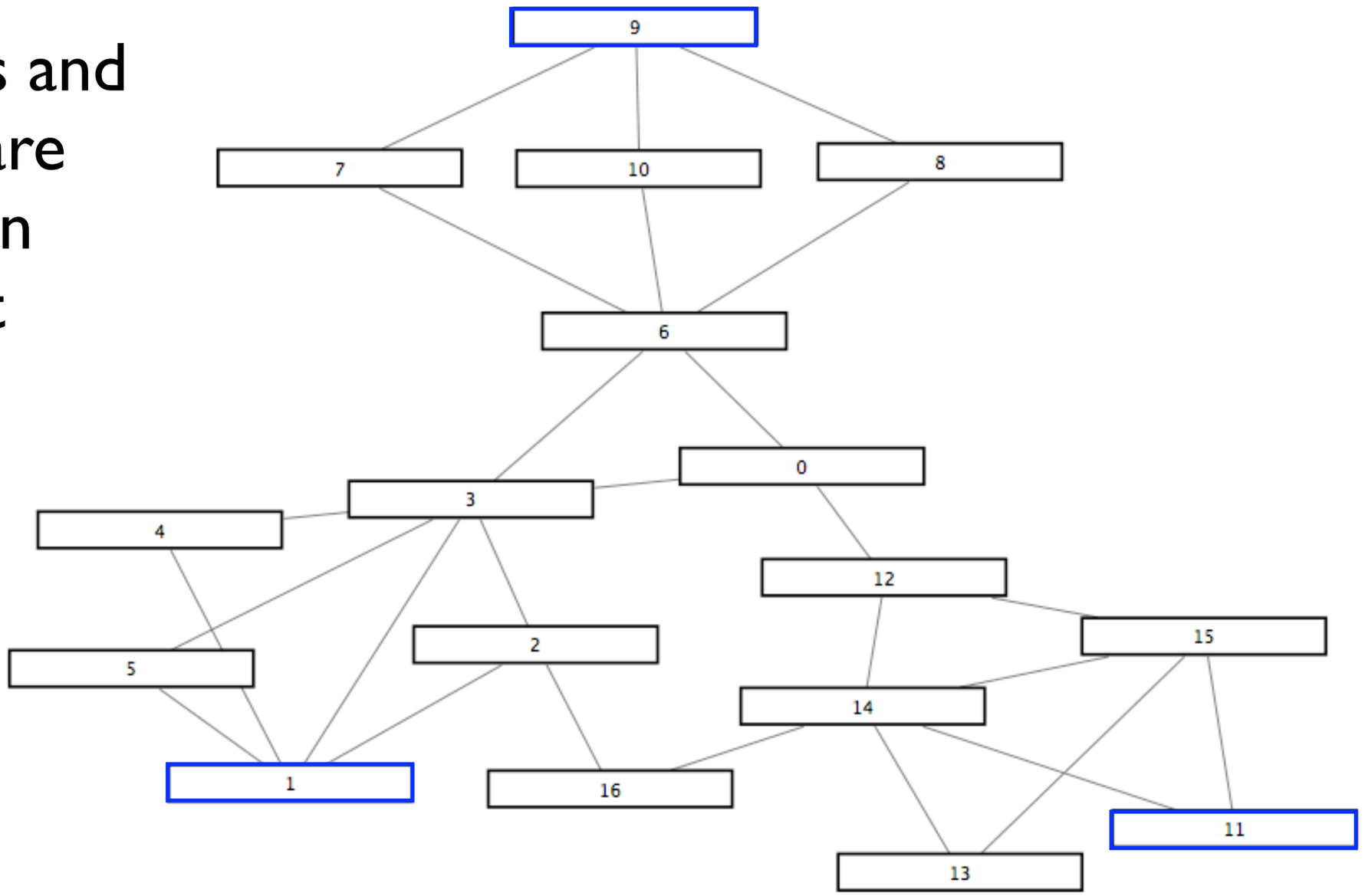
sensitivity: 1.0
positive predictive value: 0.87
arithmetic accuracy: 0.93
geometric accuracy: 0.93



3. Methods - kWalks algorithm

kWalks algorithm

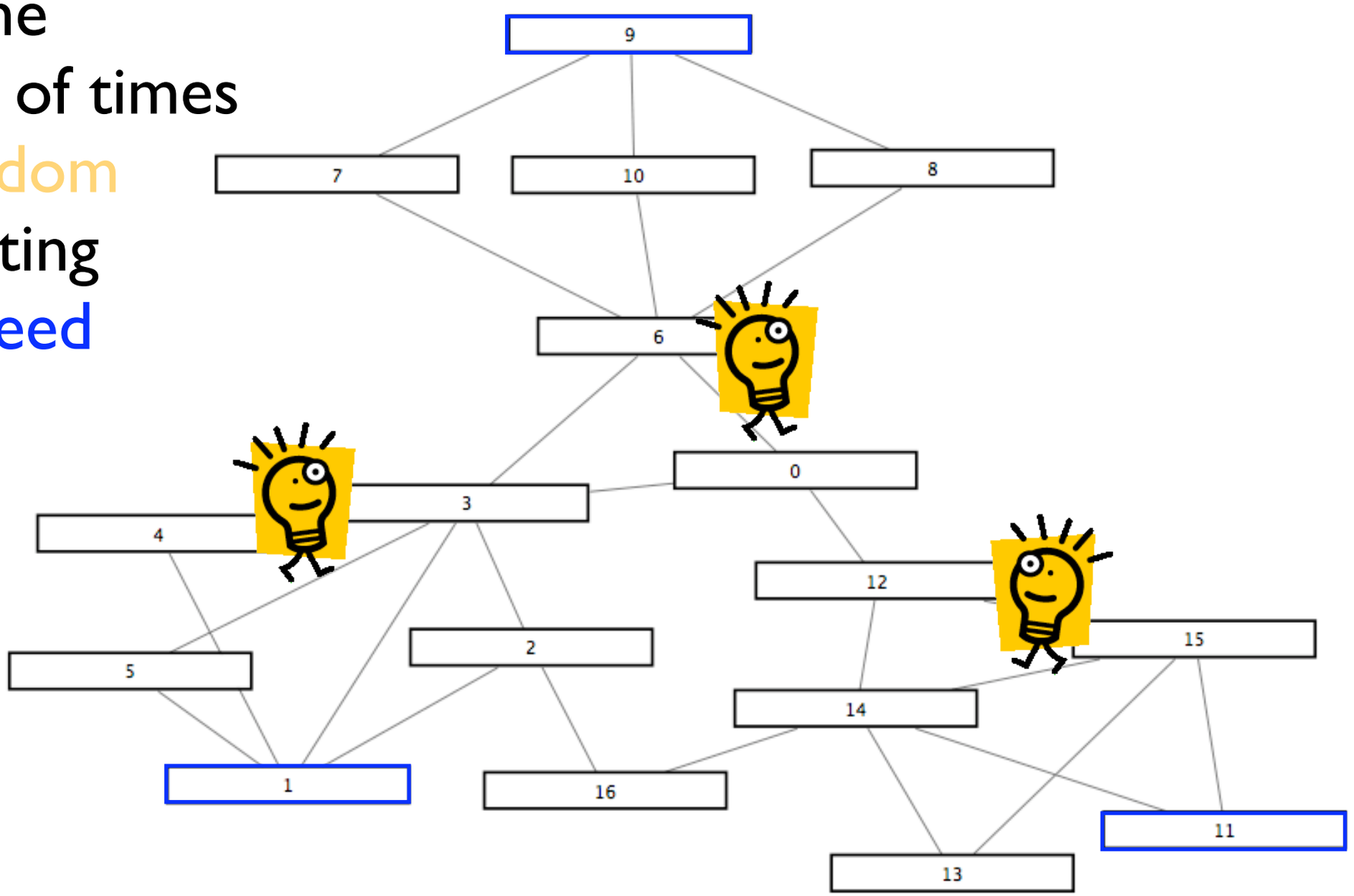
- idea: some edges and nodes in a graph are more relevant than others to connect given seed nodes



P. Dupont, J. Callut, G. Dooms, 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.

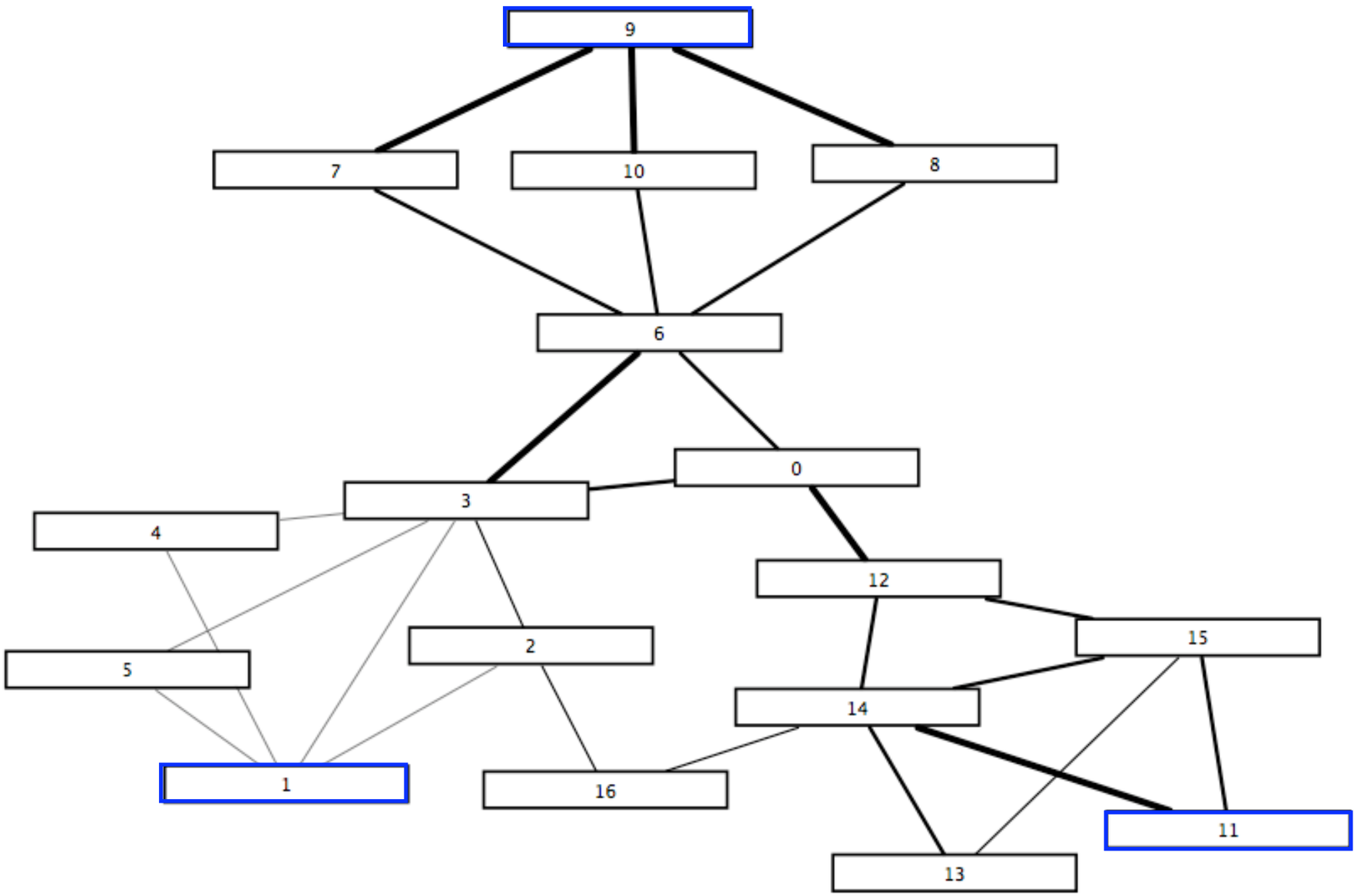
3. Methods - kWalks algorithm

- 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



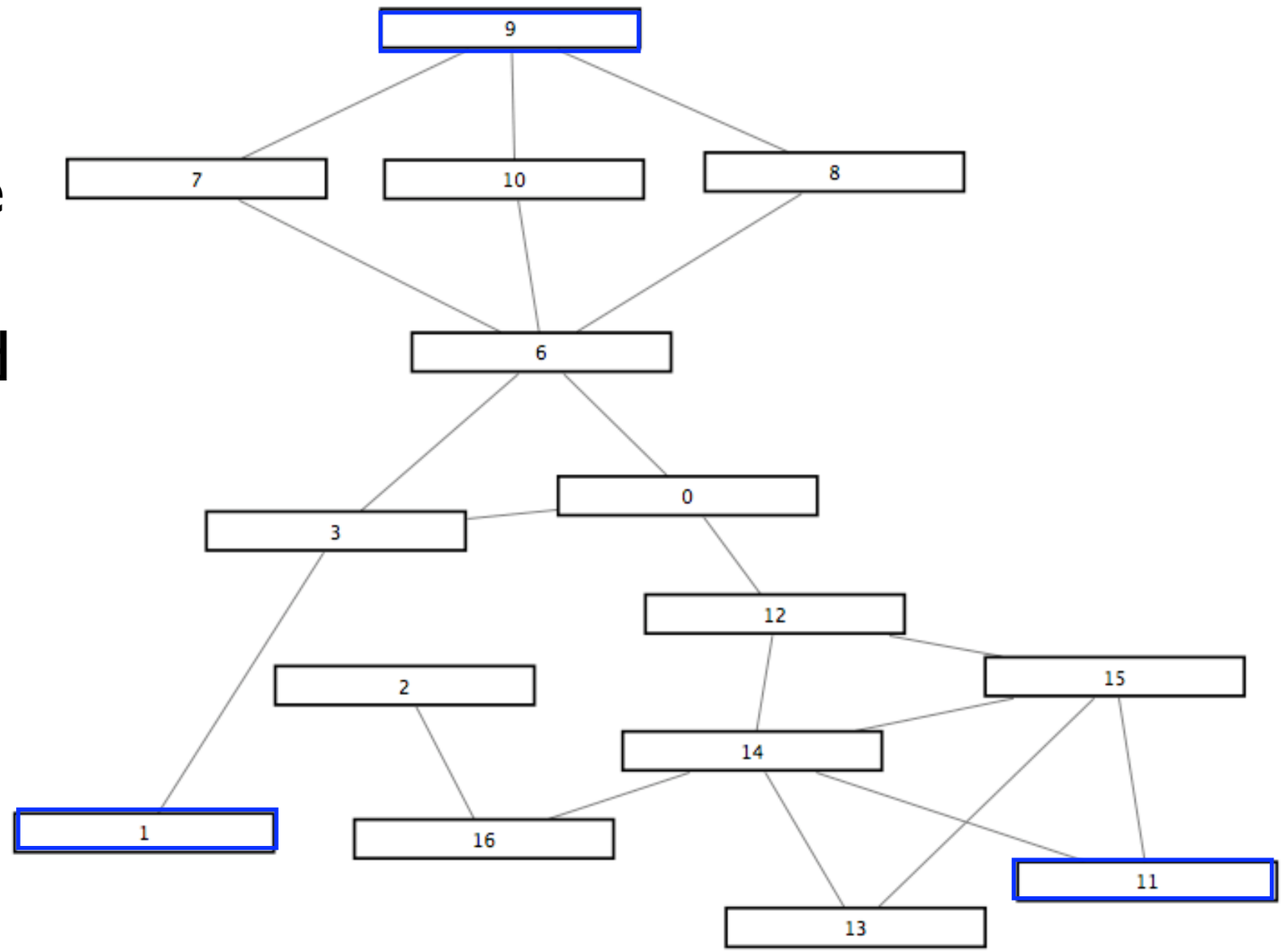
3. Methods - kWalks algorithm

- output: list of edge and node relevances



3. Methods - kWalks algorithm

- extract subgraph: add edges and their adjacent nodes in the order of their relevance to the seed nodes until seed nodes are connected

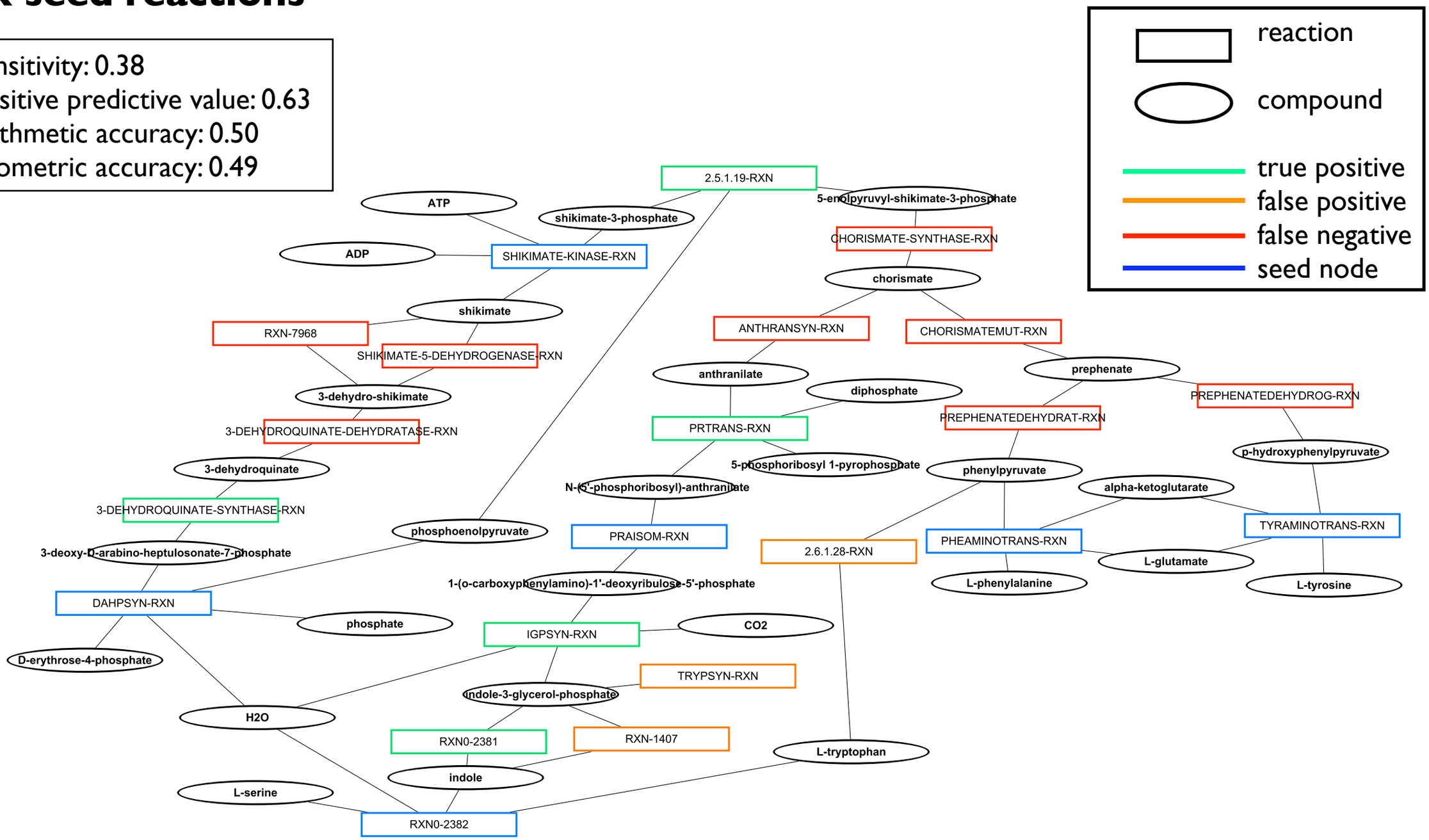


3. Methods - kWalks in unweighted MetaCyc graph

reference: superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis

six seed reactions

sensitivity: 0.38
positive predictive value: 0.63
arithmetic accuracy: 0.50
geometric accuracy: 0.49



4. Evaluation of kWalks

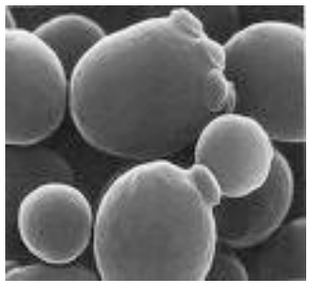
Reference pathways

- 71 pathways taken from the *Saccharomyces cerevisiae* pathways annotated in MetaCyc

- minimal pathway size: 5 nodes

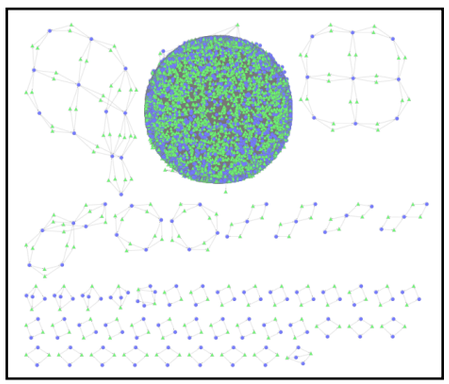
- average node number: 13

- 34 branched and 17 cyclic pathways



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

Metabolic graph



- MetaCyc (all reactions and compounds)

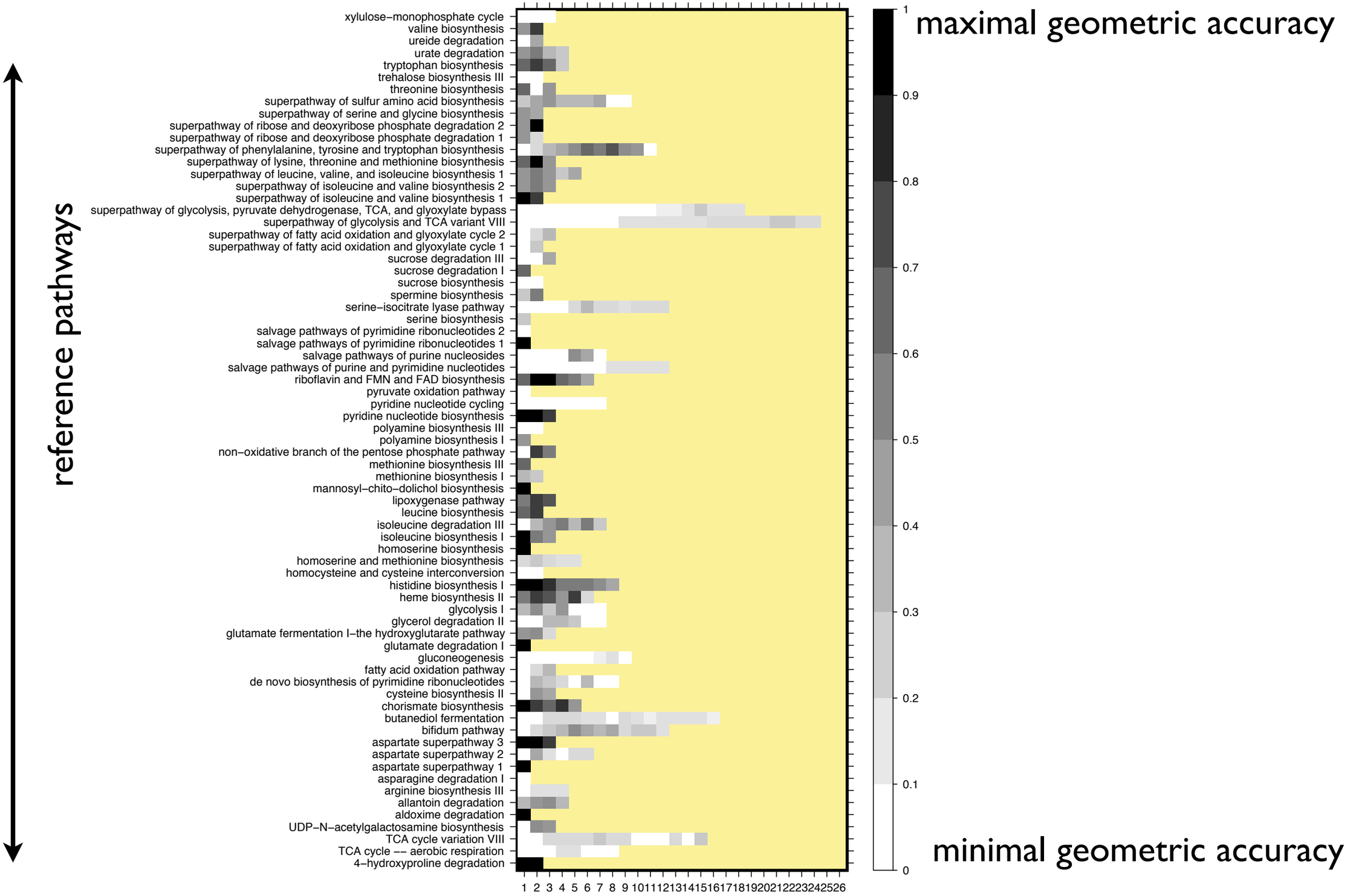
- 4,891 compound nodes and 5,358 reaction nodes

Evaluation procedure

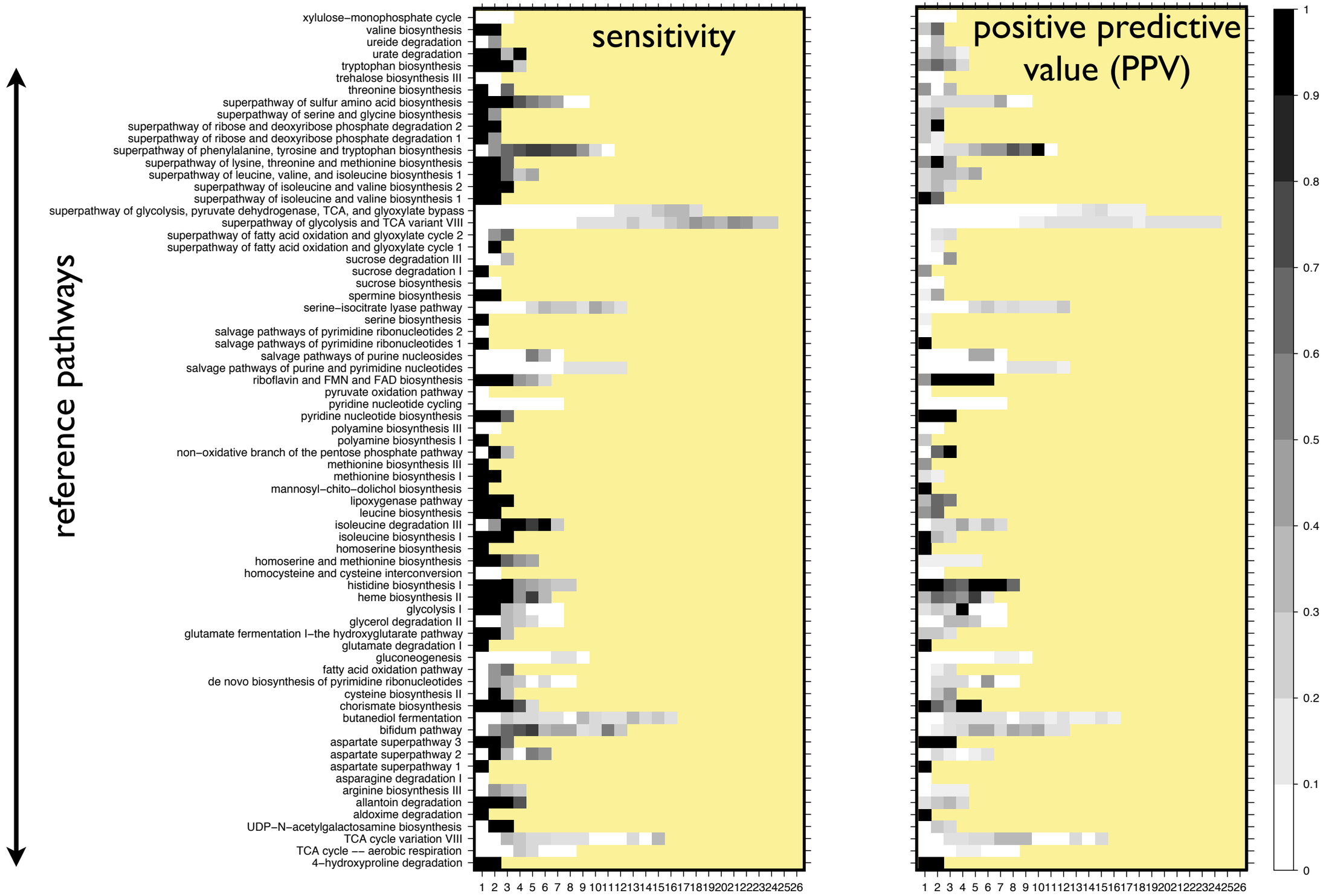
- for each reference pathway, do inference with terminal reactions of the reference pathway as seed nodes

- repeat inference by adding one additional reaction at each step to the seed reaction set

4. Evaluation of kWalks - Geometric accuracy heat map for unweighted graph



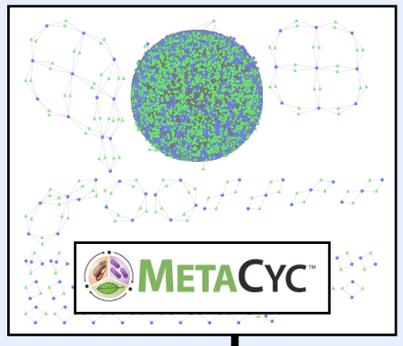
4. Evaluation of kWalks - Sensitivity and PPV heat map for unweighted graph



4. Evaluation of kWalks - Parameter optimization

Input

input graph (directed or undirected)



seed reactions

Aspartate kinase

Threonine synthase

5-methyltetrahydropteroyltriglutamate--homocysteine 5-methyltransferase

5-methyltetrahydrofolate--homocysteine 5-methyltransferase

initial edge weights

edge identifier	edge weight
PSERTRANSAM-RXN<→3-P-SERINE	0.03
2.5.1.65-RXN>→CYS	0.01
3-P-SERINE→2.5.1.65-RXN>	0.01
...	

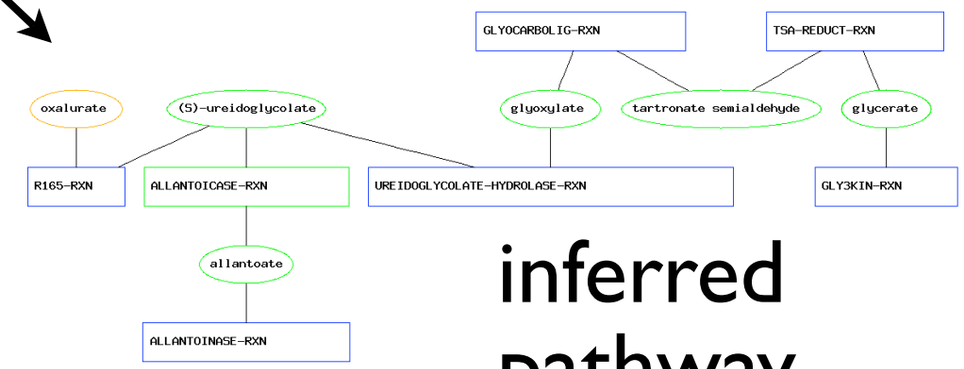
edge weights

kWalks

iteration

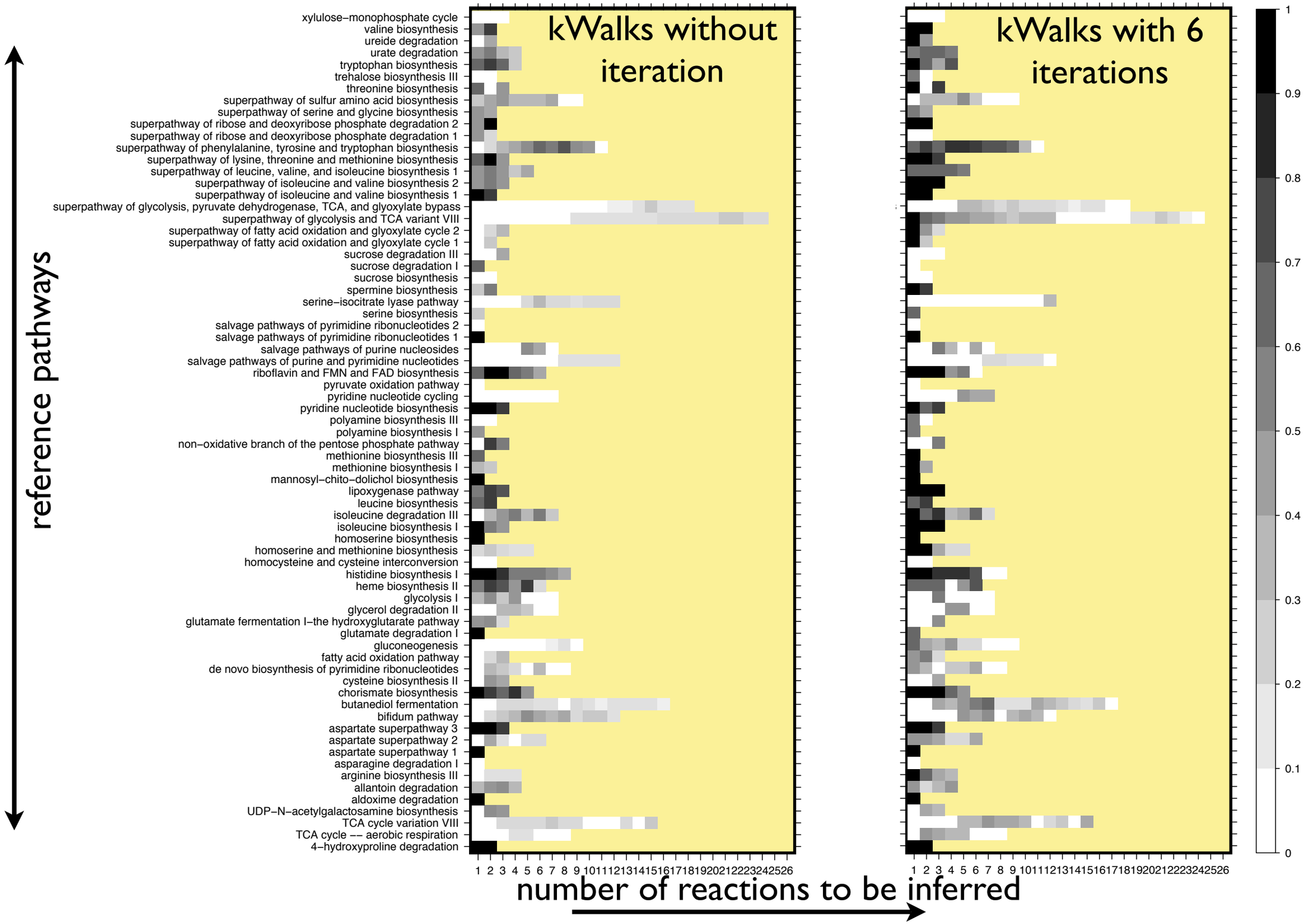
edge identifier	edge relevance
PSERTRANSAM-RXN<→3-P-SERINE	0.043
3-P-HYDROXYPYRUVATE→PSERTRANSAM-RXN<	0.043
PGLYCDEHYDROG-RXN>→3-P-HYDROXYPYRUVATE	0.041
2.5.1.65-RXN>→CYS	0.026
3-P-SERINE→2.5.1.65-RXN>	0.026

edge relevances



inferred pathway

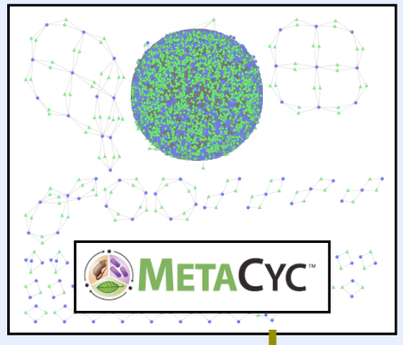
4. Evaluation of kWalks - Geometric accuracy heat map without and with iteration



4. Evaluation of kWalks - Parameter optimization

Input

input graph (directed or undirected)



seed reactions

Aspartate kinase

Threonine synthase

5-methyltetrahydropteroyltriglutamate--homocysteine S-methyltransferase

5-methyltetrahydrofolate--homocysteine S-methyltransferase

initial edge weights

edge identifier	edge weight
PSERTRANSAM-RXN<→3-P-SERINE	0.03
2.5.1.65-RXN>→CYS	0.01
3-P-SERINE→2.5.1.65-RXN>	0.01
...	

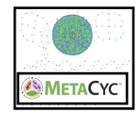
edge weights

kWalks

iteration

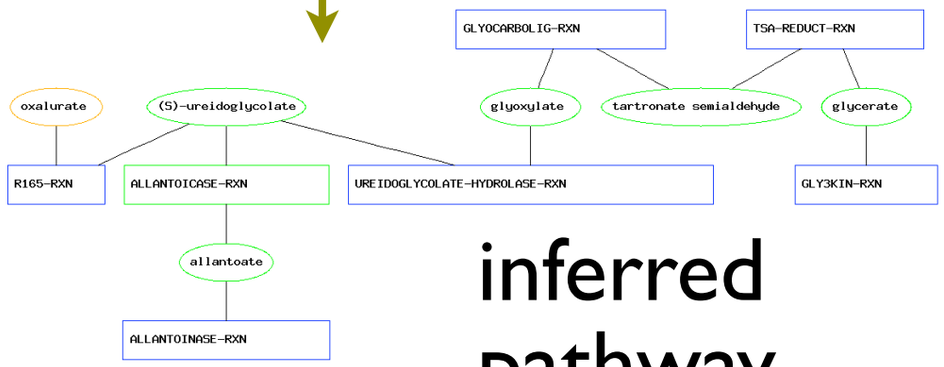
edge identifier	edge relevance
PSERTRANSAM-RXN<→3-P-SERINE	0.043
3-P-HYDROXYPYRUVATE→PSERTRANSAM-RXN<	0.043
PGLYCDEHYDROG-RXN>→3-P-HYDROXYPYRUVATE	0.041
2.5.1.65-RXN>→CYS	0.026
3-P-SERINE→2.5.1.65-RXN>	0.026

edge relevances



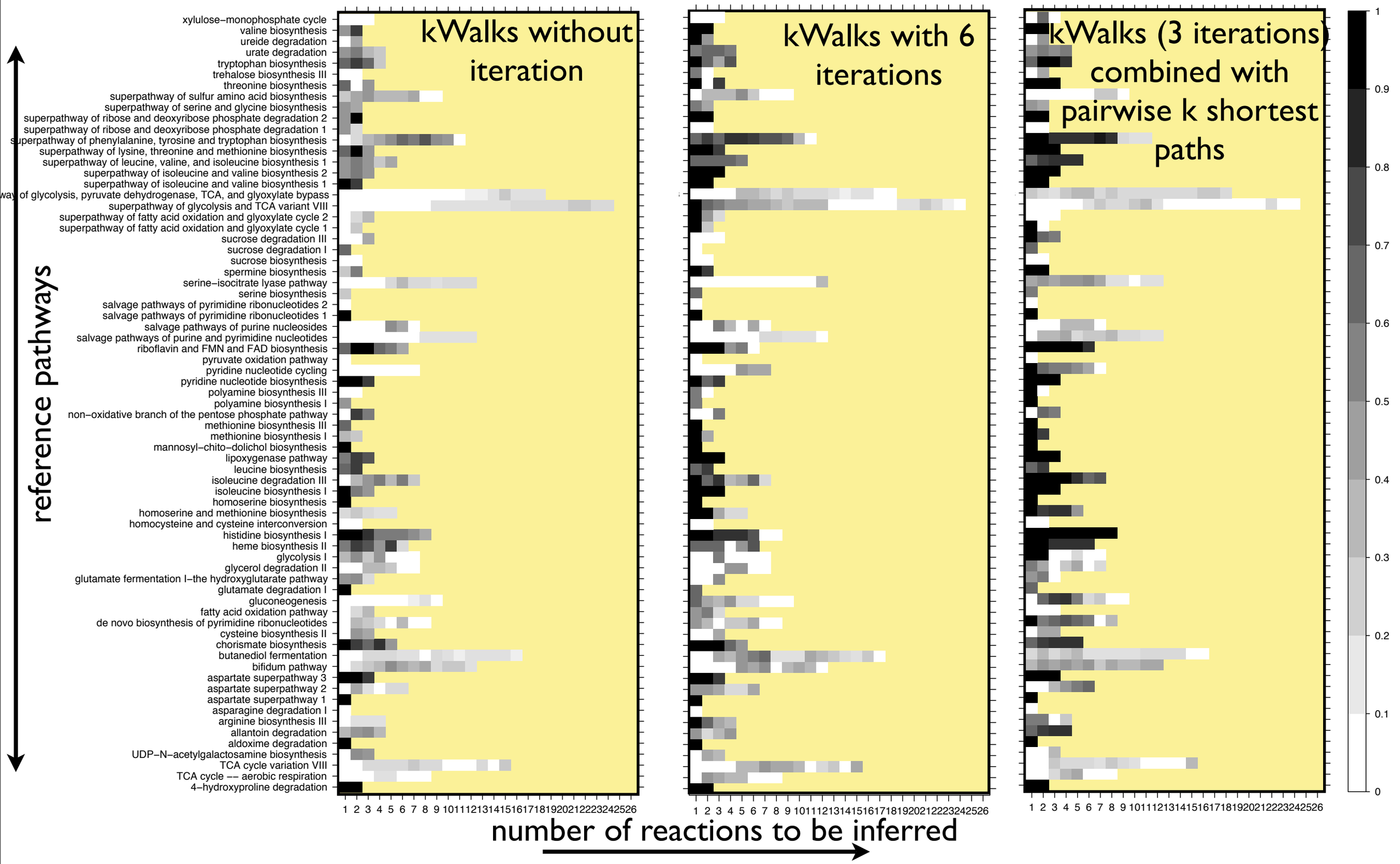
input graph of reduced size

pair-wise k shortest paths



inferred pathway

4. Evaluation of kWalks - Accuracy heat map without and with pairwise k shortest paths



4. Evaluation of kWalks - Summary

- kWalks is much faster (order of seconds) than pair-wise k shortest paths (order of minutes)

- iterating kWalks or combining it with pair-wise k shortest paths reduces number of false positives

- in contrast to pair-wise k shortest paths, kWalks avoids hub nodes in unweighted graphs

- kWalks performs slightly better in directed than in undirected MetaCyc graph

5. Conclusion

- kWalks and pairwise k shortest paths complementary:
    - kWalks: high sensitivity, quick
    - pairwise k shortest paths: high positive predictive value for a high computational cost
    - combination of both: promising approach for pathway inference in metabolic graphs

6. Next Steps

- test Steiner tree algorithms in combination with kWalks

- improve pathway inference by considering main/side compound annotation (work in progress)

- test approach on microarray data

- make pathway inference available as Web Service

Acknowledgement

ULB

BiGRe
Bioinformatique des Génomes et Réseaux

Jacques van Helden (supervisor)
Didier Croes
BiGRe team

UCL

INGI

machine learning group @ UCL

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

IBMM
Bruno André
Patrice Godard

aMAZE

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

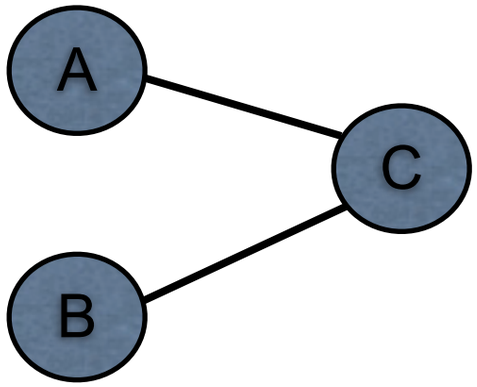
The PhD grant of Karoline Faust is funded by the Actions de Recherche Concertées de la Communaute Française de Belgique (ARC grant number 04/09-307). The INGI-SCMBB collaboration is funded by the Région Wallonne de Belgique (projects aMAZE and TransMaze).

Appendix I - Graph representation of metabolic data

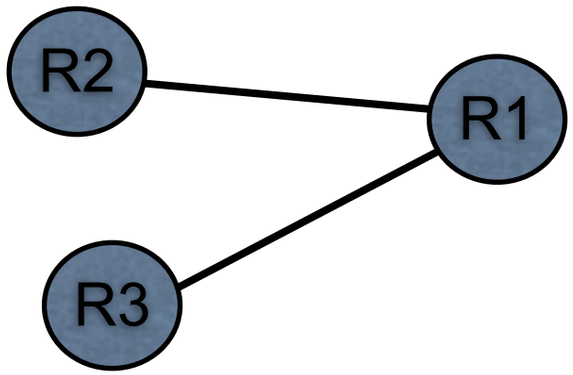
Why bipartite?

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

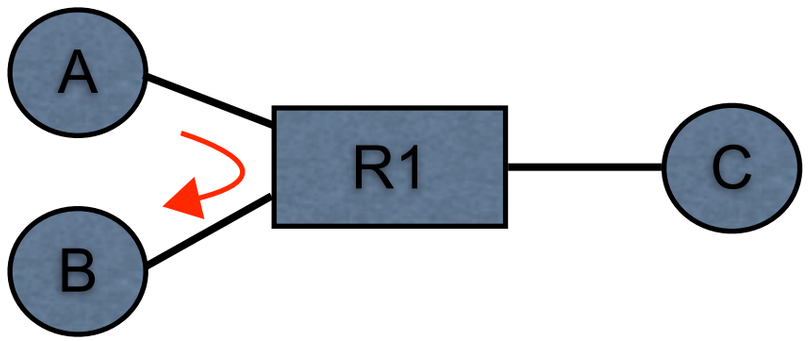


compound A is represented by several edges

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

J. van Helden, L. Wernisch, D. Gilbert, S. Wodak, “Graph-based analysis of metabolic networks”, Ernst Schering Research Foundation Workshop, Springer-Verlag 38 (2002), 245-274.

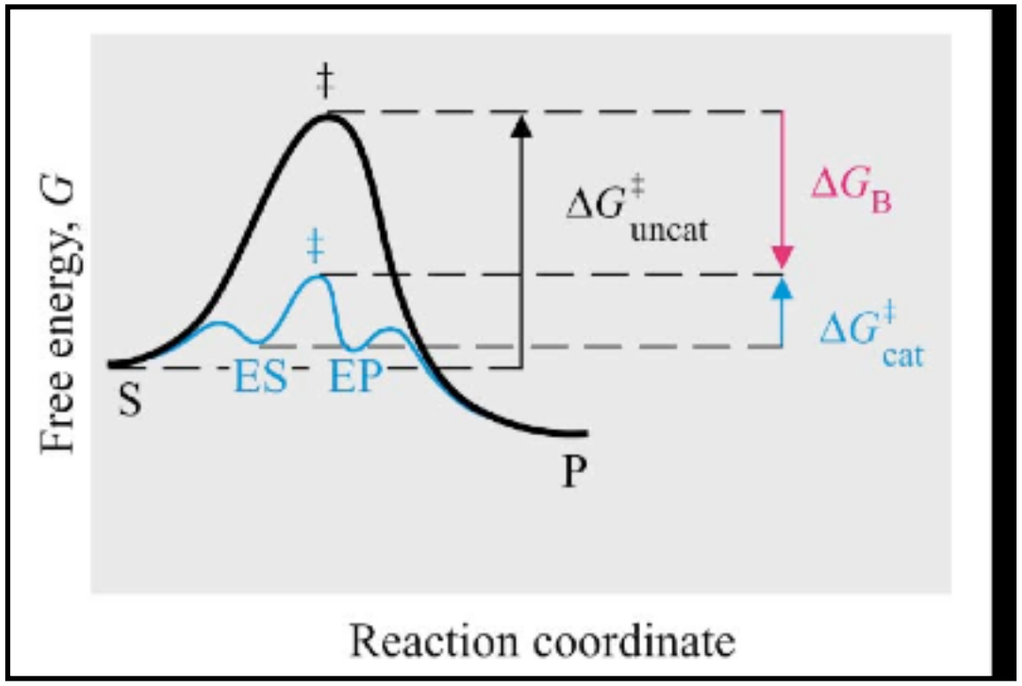
Appendix II - Graph representation of metabolic data - directionality

- 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 educt 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 educt and product concentrations, instead they speed up attainment of equilibria:



$\Delta G = \Delta G^{\circ} + RT \ln([product_1]...[product_m]/[educt_1]...[educt_n])$

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

Appendix III - MetaCyc graph

Parsing

- from MetaCyc (Release 11.0) owl file (MetaCyc: collection of well annotated organisms in BioCyc)

- restriction to small molecule compounds and reactions having as educts/products small molecules (graph represents small molecule metabolism)

Processing

- removal of orphan nodes

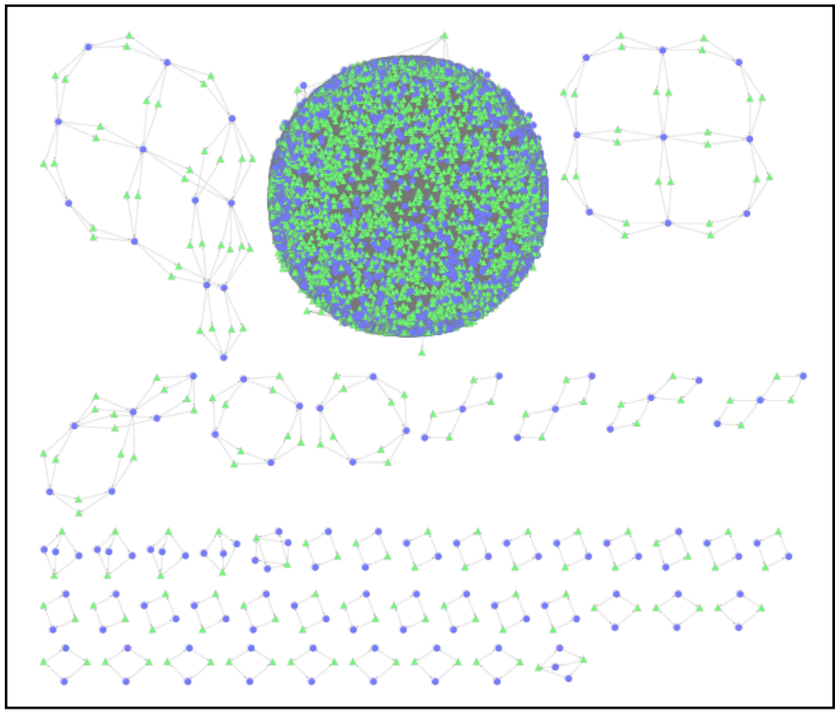
- removal of reactions having the same compound as educt and product

Properties

- 4,891 compound nodes and 5,358 reaction nodes

- 43,938 arcs

- 52 strongly connected components



Appendix IV - Reference pathways

Parsing

- 171 pathways obtained from BioCyc (Release 11.0) *S. cerevisiae* owl file

- side/main compound annotation obtained from *S. cerevisiae* pathway.dat file

Processing

- removal of pathways with node identifiers absent from the largest strongly connected component of the MetaCyc graph

- removal of pathways with less than five nodes (inference would be trivial)

- after processing, 71 pathways left

Appendix V - 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 kWalks

- 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:

$$w(a)^z$$

Appendix VI - Metabolic path finding evaluation

- Validation of metabolic path finding with KEGG/LIGAND graph and metabolic pathways annotated in aMAZE database

Shortest path			
Graph	**Average sensitivity**	**Average PPV**	**Average accuracy**
Raw	31.4%	25.4%	**28.4%**
Filtered	68.0%	63.0%	**65.5%**
Weighted	88.5%	83.4%	**85.9%**

- Validation of metabolic path finding with EcoCyc graph and metabolic pathways annotated in EcoCyc

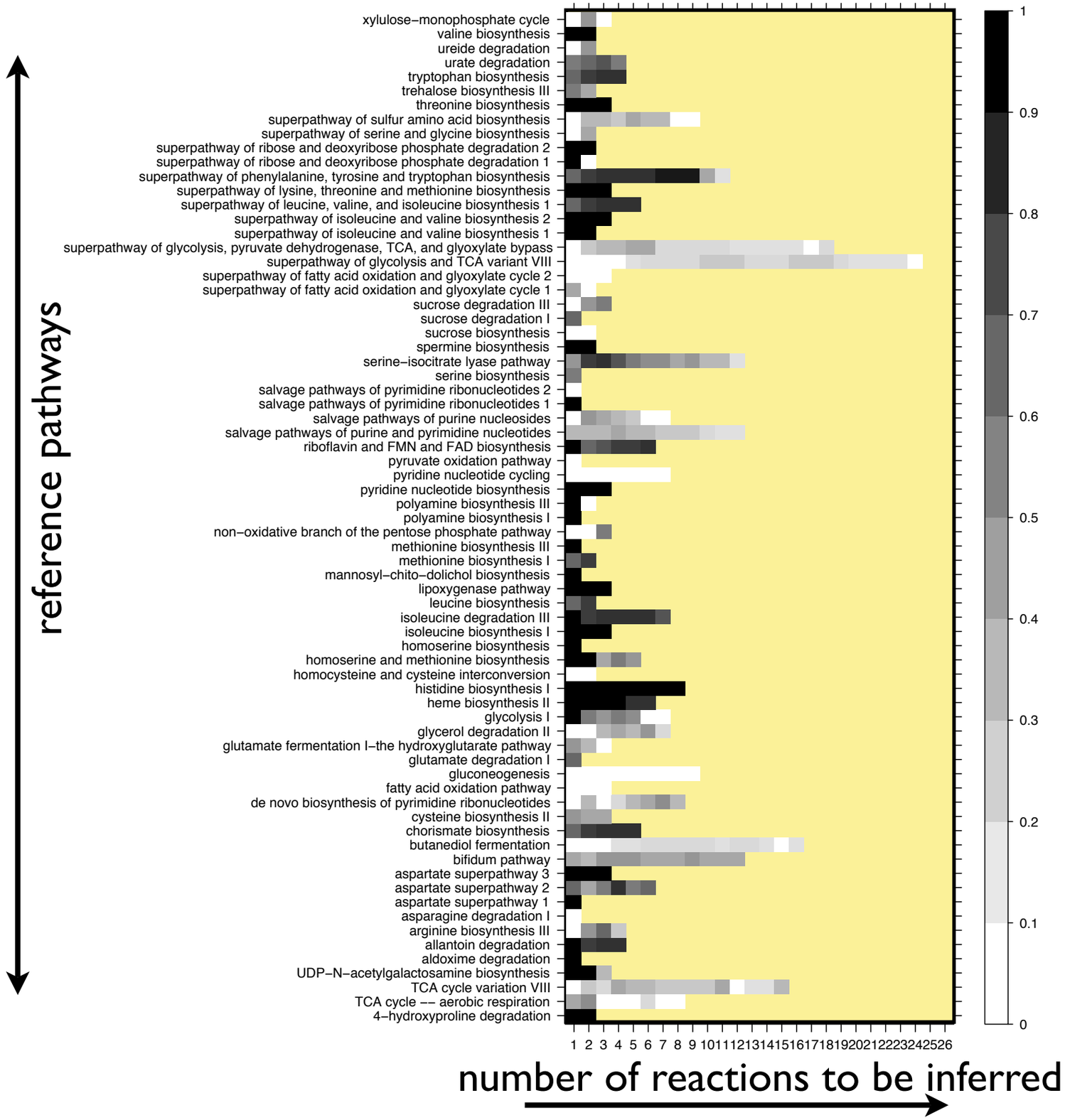
Shortest path			
Graph	**Average sensitivity**	**Average PPV**	**Average accuracy**
Raw	29.6%	31.0%	**29.3%**
Filtered	63.3%	68.8%	**66.6%**
Weighted	80.7%	85.3%	**83.0%**

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.

C. Lemer, H. Anerhour, J.M. Maniraja, O. Sand, J. Richelle and S. Wodak (2004). “The aMAZE database goes public.” ECCB.

Appendix VII - Evaluation pairwise k shortest paths in weighted MetaCyc graph



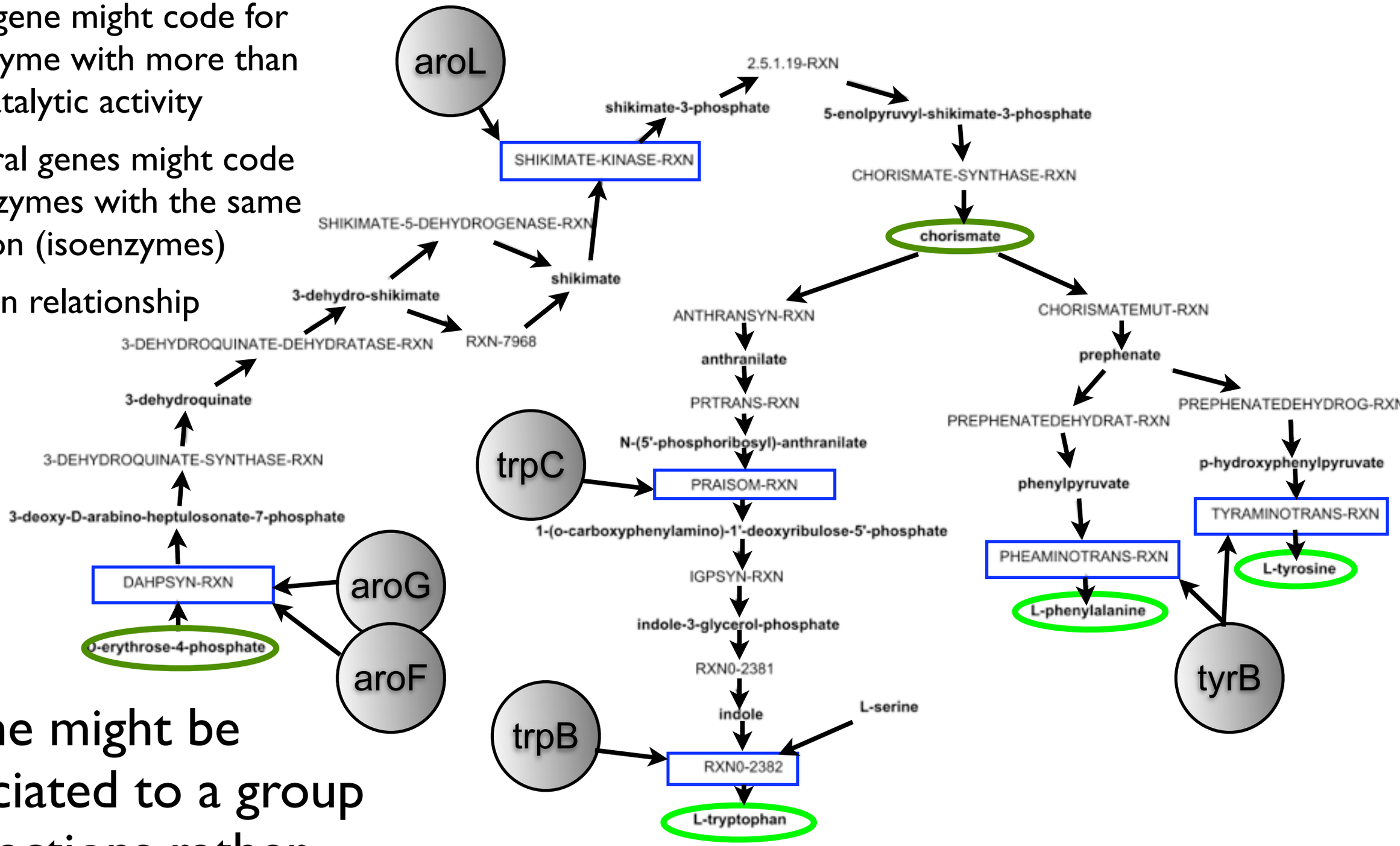
geometric accuracy heat map

Appendix VIII - Gene to reaction mapping

- one gene might code for an enzyme with more than one catalytic activity

- several genes might code for enzymes with the same function (isoenzymes)

- n-to-n relationship



a gene might be associated to a group of reactions rather then one reaction

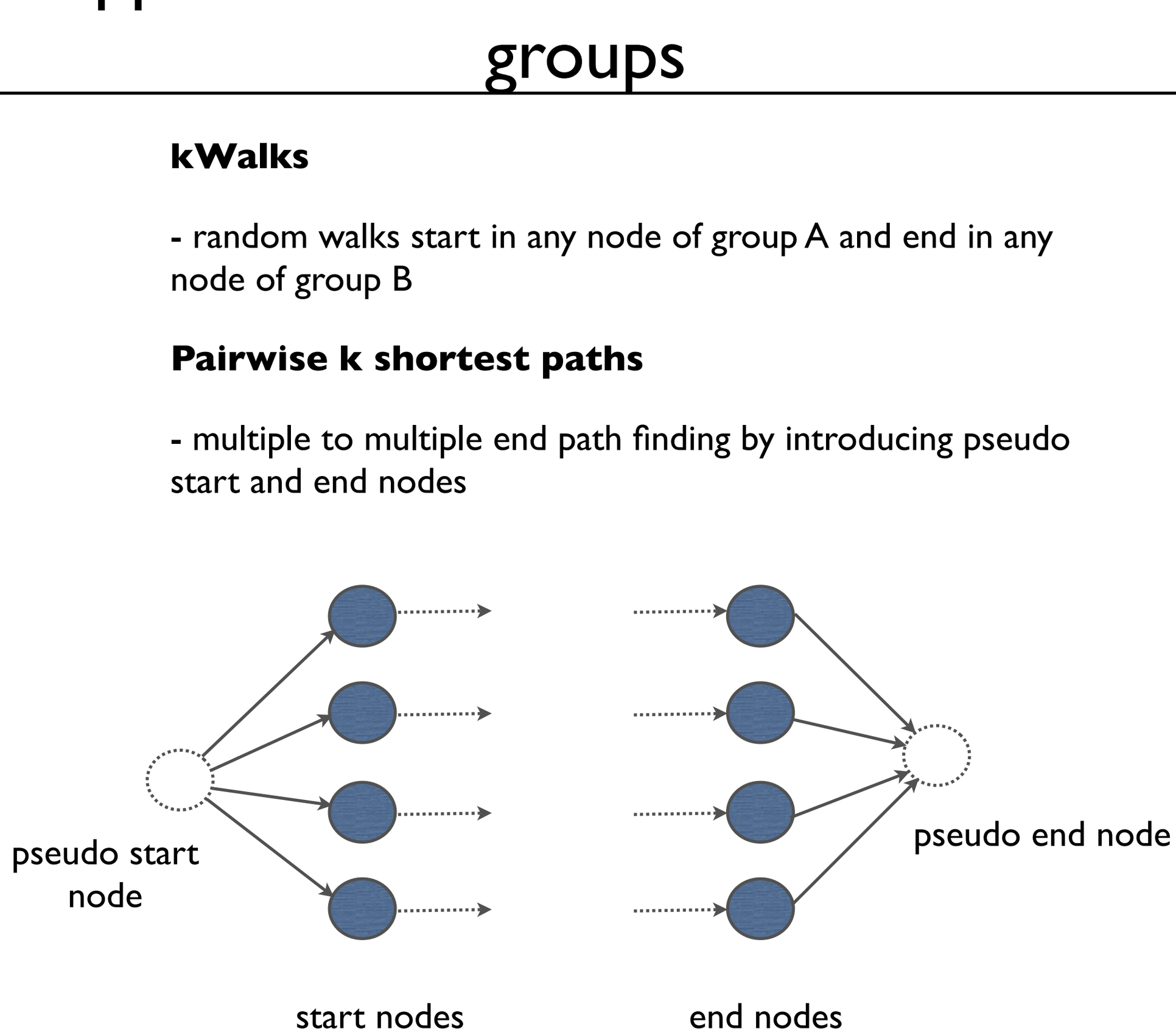
Appendix IX - Treatment of reaction groups

kWalks

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

Pairwise k shortest paths

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



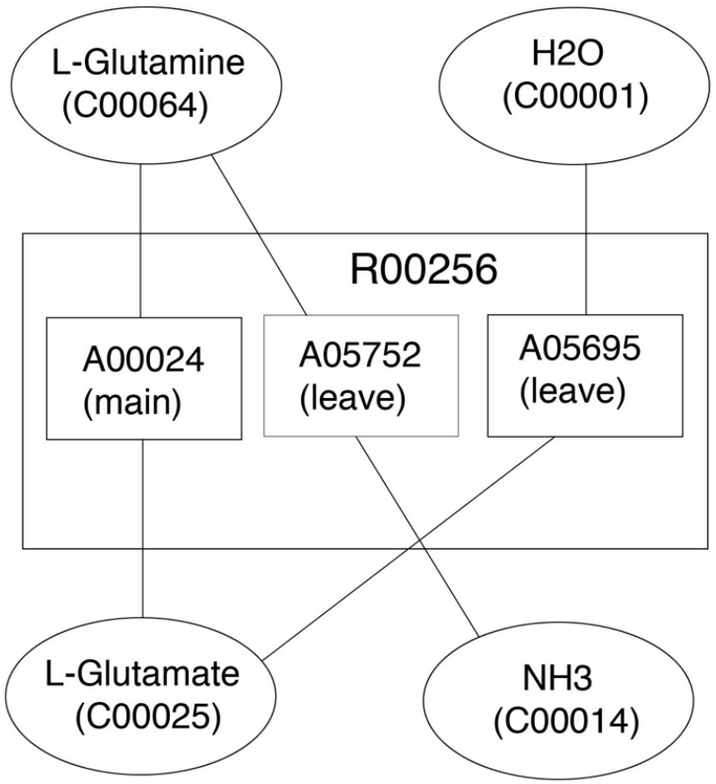
Appendix X - Main/side compounds

Basic idea

- main/side compound annotation present in KEGG/LIGAND in form of sub-reactions (RPairs)

- favor sub-reactions that connect main compounds

sub-reactions



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.