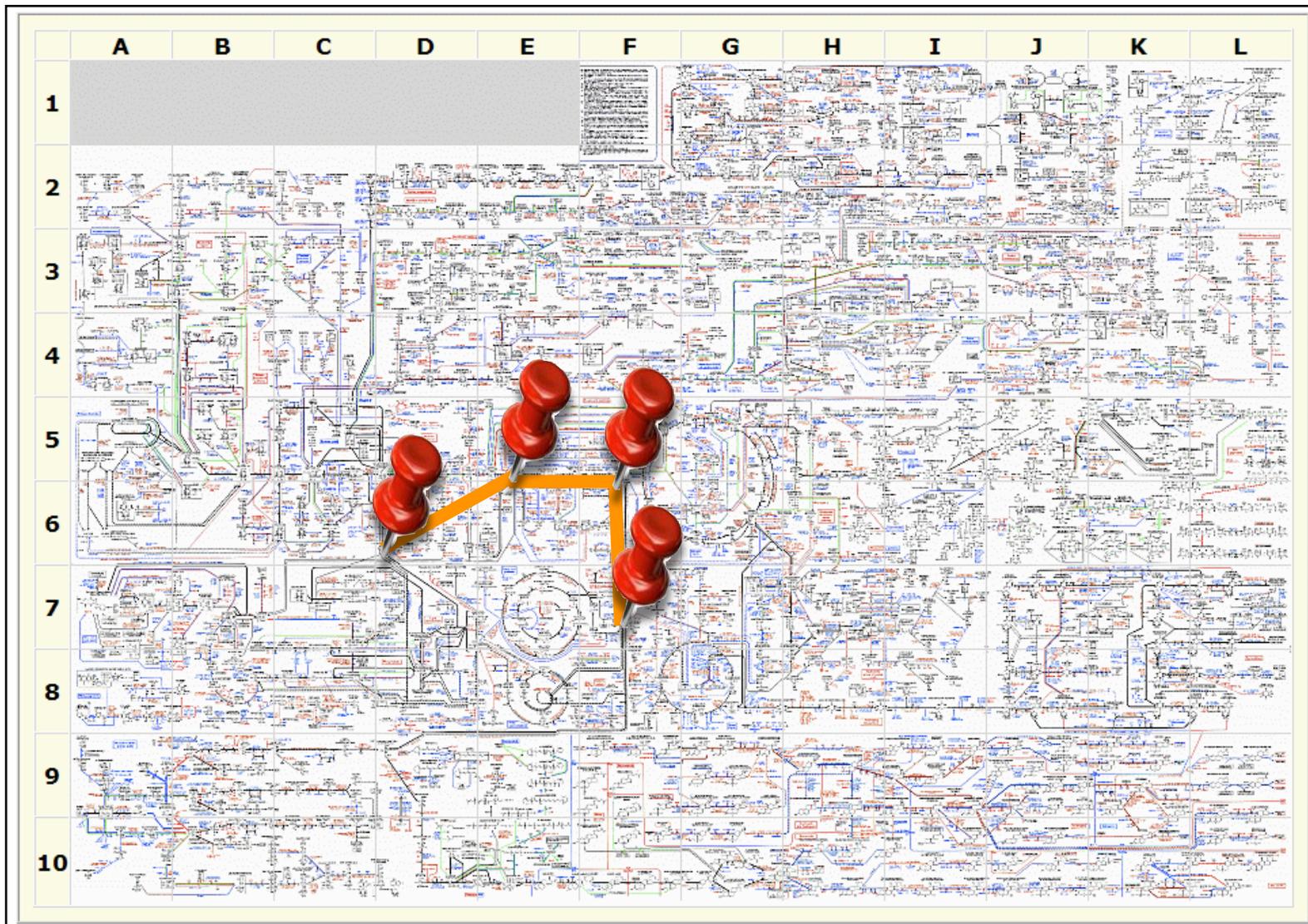


Predicting metabolic pathways from functionally linked genes

Karoline Faust, Didier Croes and Jacques van Helden



Functionally linked genes

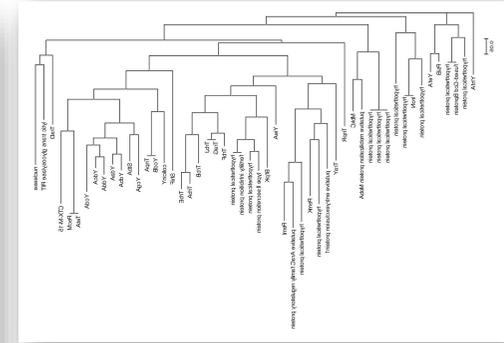
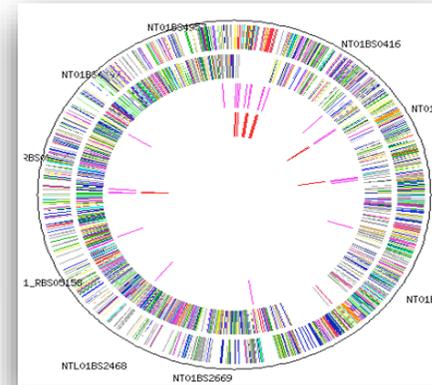
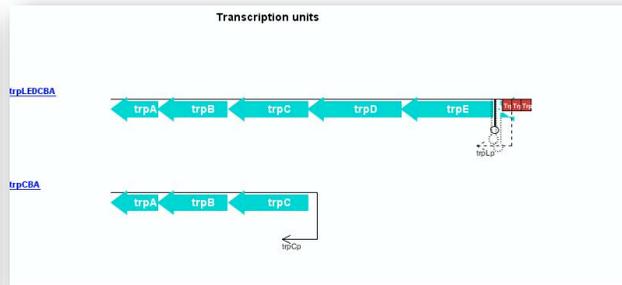
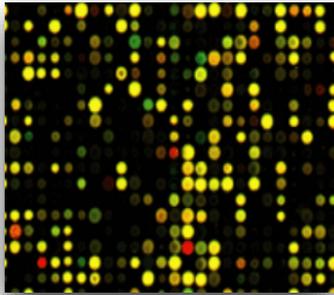


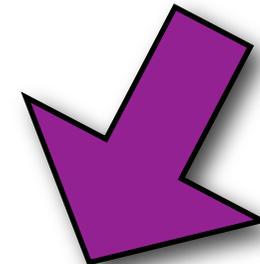
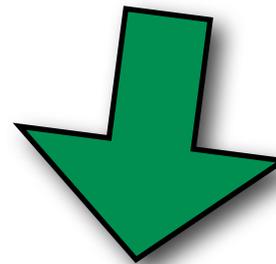
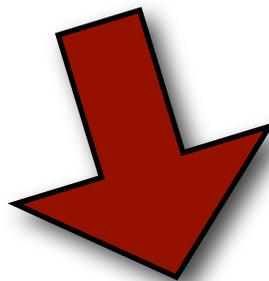
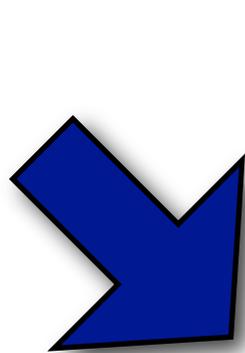
Image sources: University of Liverpool Microarray facility, RegulonDB, Comprehensive Microbial Resources, Brilli *et al. BMC Bioinformatics* 2008 **9**:551

co-expression
of genes in
microarray
data

co-regulation of
genes in **operons**
and **regulons**

co-localization of
genes in the
genome

co-occurrence of
genes across taxa
(phylogenetic
profiles)



groups of associated genes
assumed to be involved in a
common function

Pathway mapping/Pathway projection

- common approach to link genes to functions: map genes to known functional units (reference pathways)
- example: map genes of operon aruCFGDB (*Pseudomonas aeruginosa* PAO1)
- pathway mapping tool: KEGG Mapper

PA0895 (argD)

PA0896 (aruF)

PA0897 (aruG)

PA0898 (astD)

PA0899 (aruB)

Pathway Search Result

Sort by the pathway list

Show all objects

- pae00330 Arginine and proline metabolism - *Pseudomonas aeruginosa* PAO1 (5)
- pae00300 Lysine biosynthesis - *Pseudomonas aeruginosa* PAO1 (1)
- pae01110 Biosynthesis of secondary metabolites - *Pseudomonas aeruginosa* PAO1 (1)
- pae01100 Metabolic pathways - *Pseudomonas aeruginosa* PAO1 (1)

Problems of pathway mapping



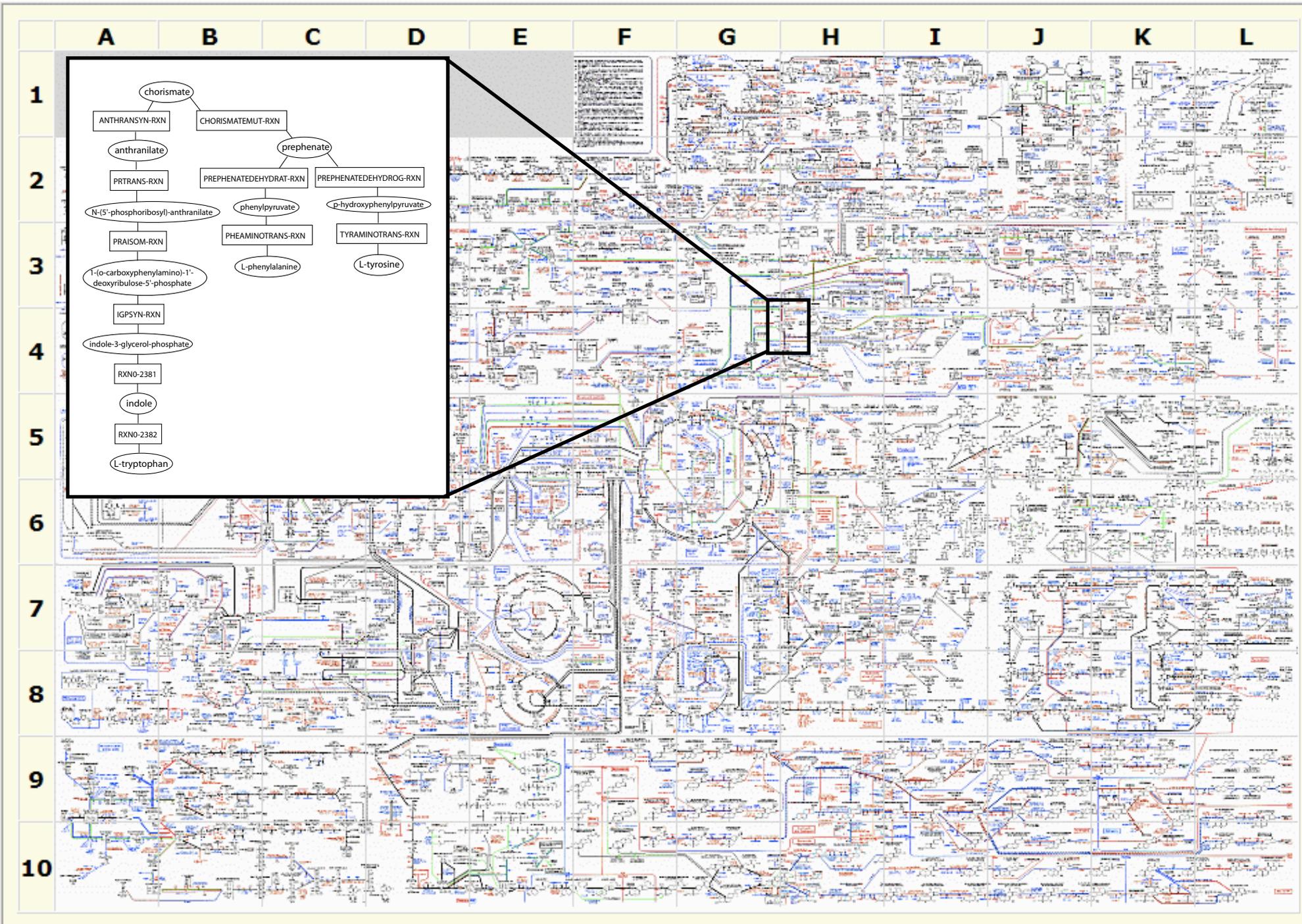
- mapping does not deal well with query genes hitting multiple reference pathways
- it cannot detect organism-specific variants of known pathways
- it cannot discover novel pathways composed of known building blocks



6 lysine biosynthesis variants listed in MetaCyc's pathway ontology

De novo discovery of metabolic pathways

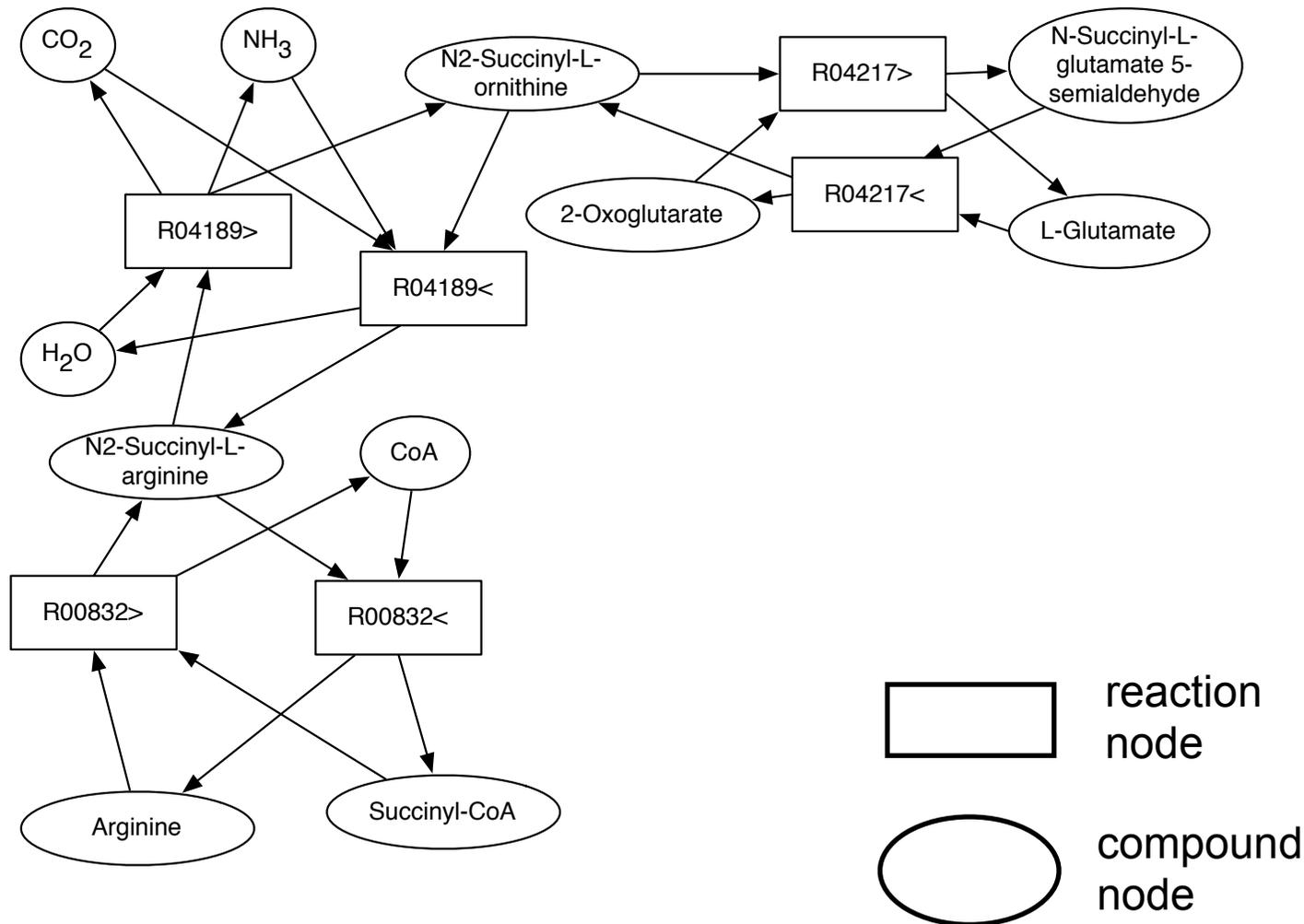
1. Introduction



Digitized version of the Roche Applied Science "Biochemical Pathways" wall chart.

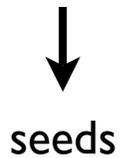
Network representation of metabolism

- metabolic network represented as **weighted** bipartite graph with two node sets: a **compound node** set and a **reaction node** set
- nodes are connected by directed **arcs**



Metabolic pathway prediction approach

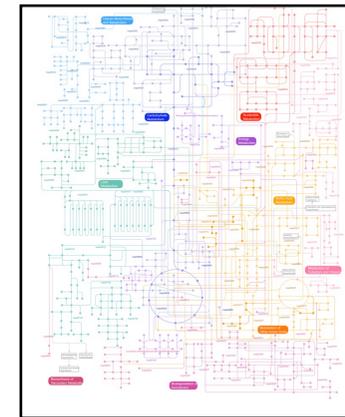
genes, enzymes,
reactions, compounds



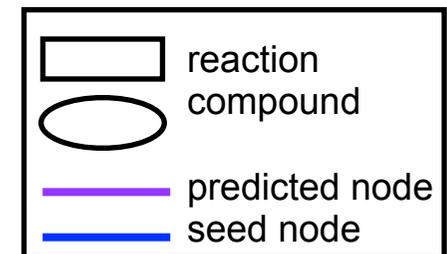
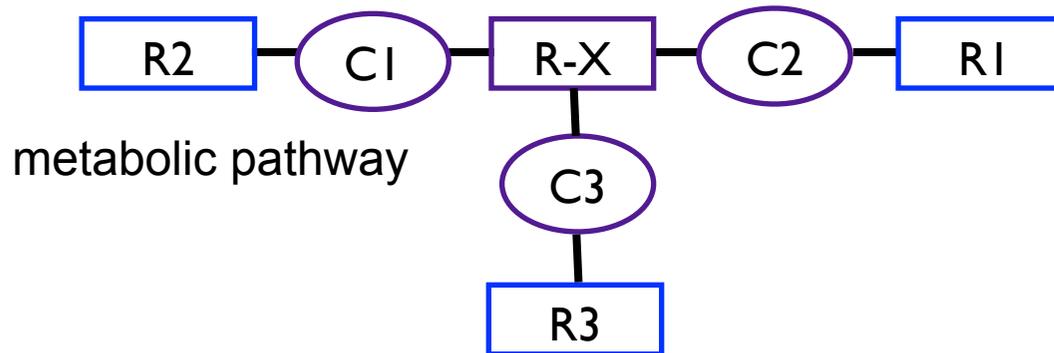
metabolic databases



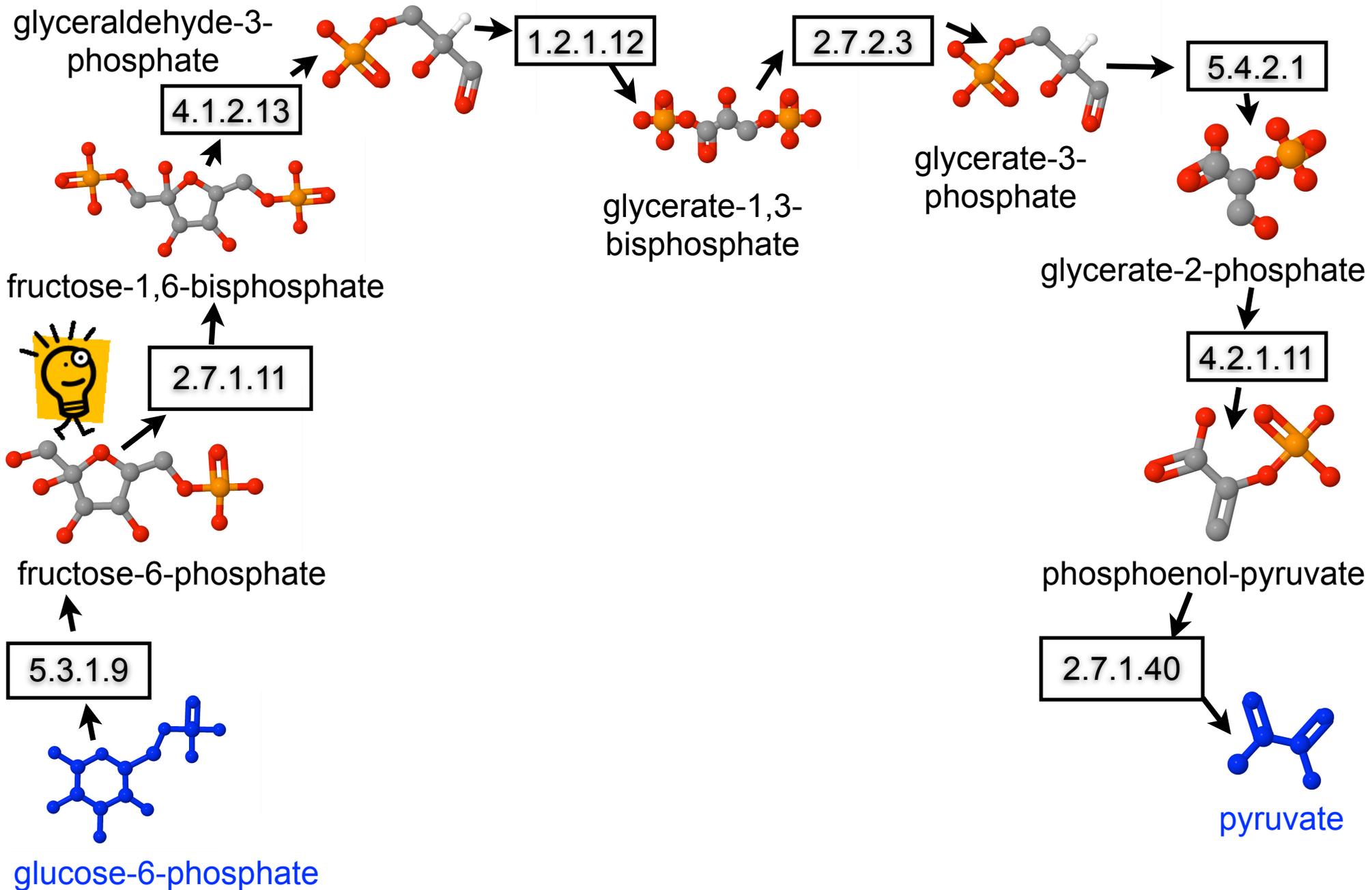
↓
metabolic network (graph)



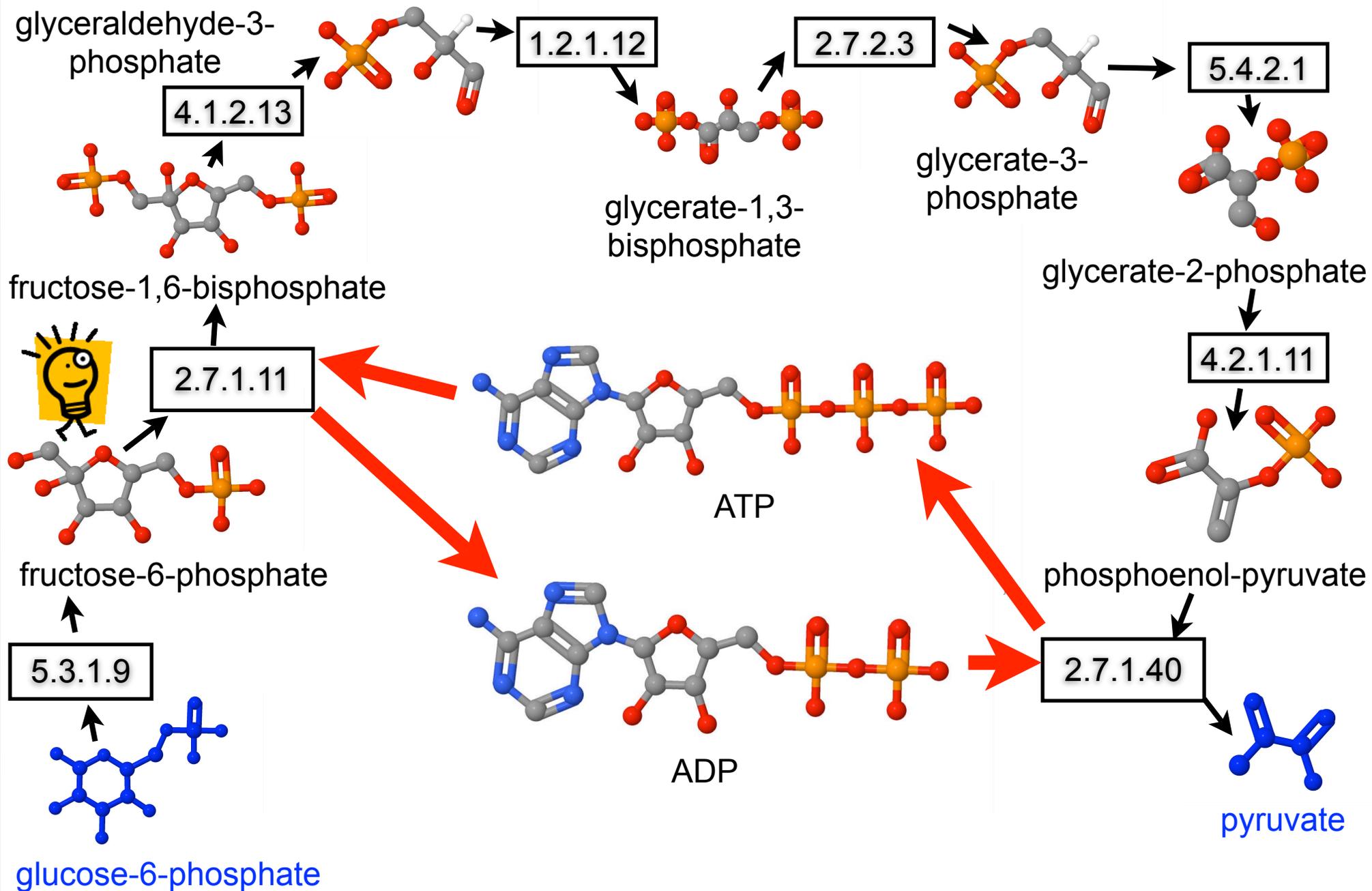
sub-network
extraction



Hub compound problem in pathway prediction



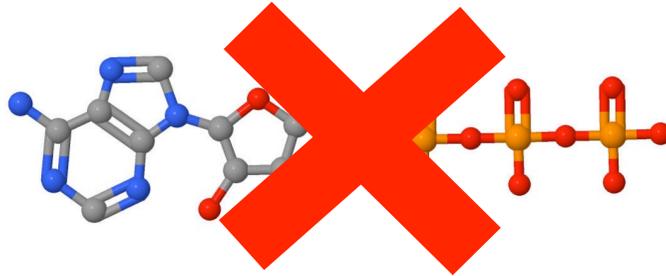
Hub compound problem in pathway prediction



2. Methods

shortcut via ADP results in biochemically invalid pathway

Hub compound problem: Solution 1



remove hub compounds from the network

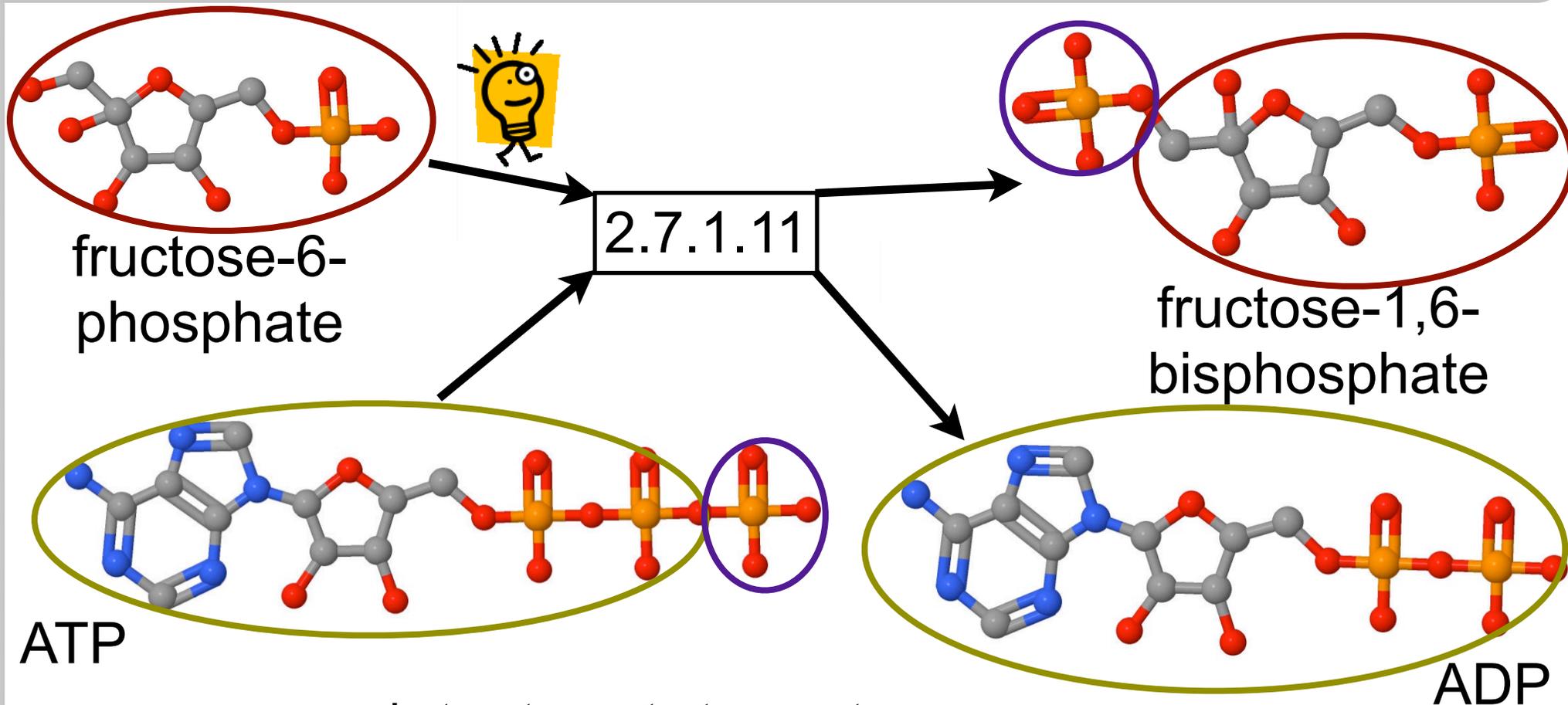
Problem 1: Which are the hub compounds?

Problem 2: What about pathways that do contain hub compounds (e.g. ATP biosynthesis)?

J. van Helden, L. Wernisch, D. Gilbert and S. Wodak (2002). "Graph-based analysis of metabolic networks." Ernst Schering Res Found Workshop, 38:245-274.

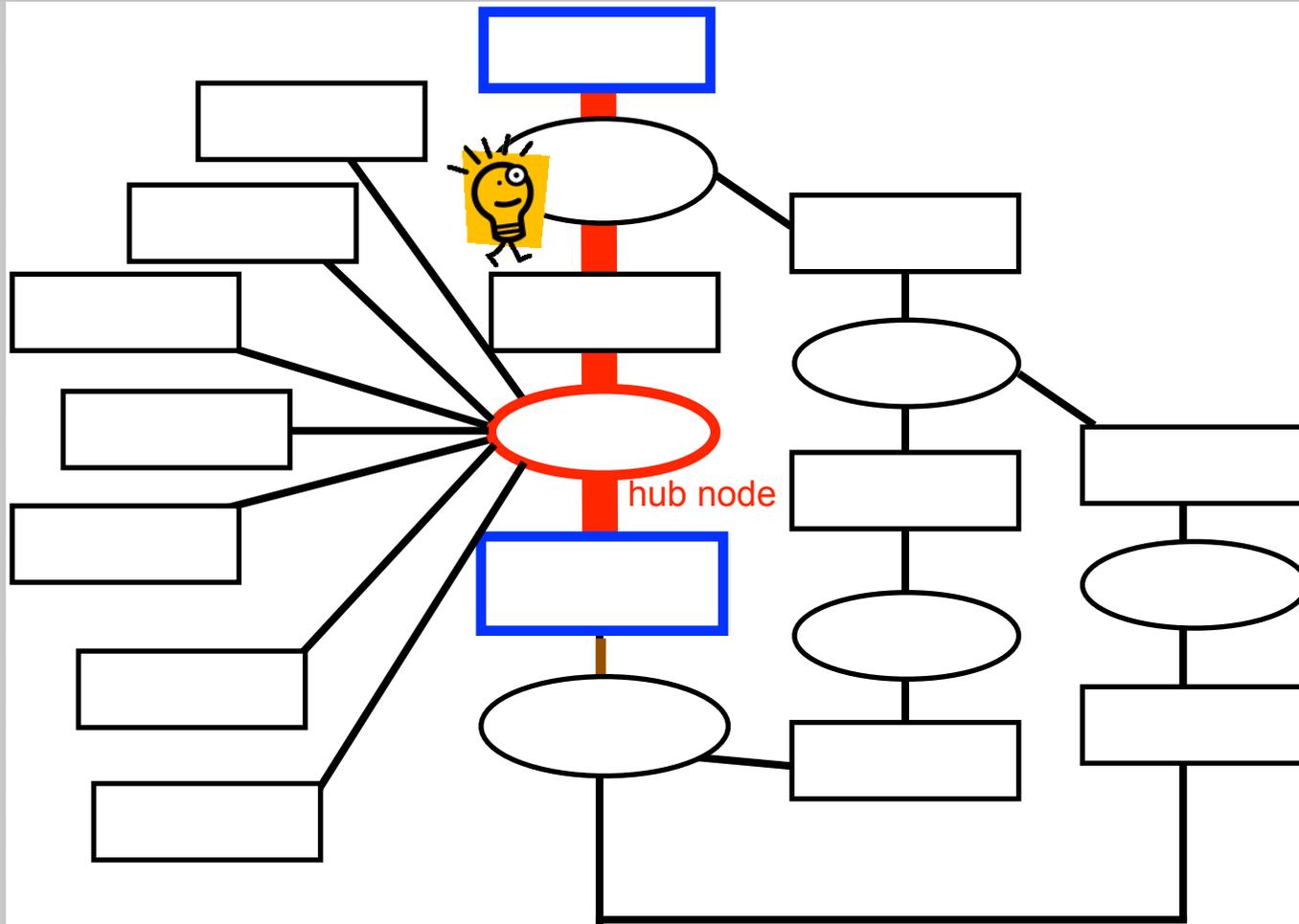
D.A. Fell and A. Wagner (2000). "The small world of metabolism." Nature, 18:1121-1122.

Hub compound problem: Solution 2



- use compound structures to trace atoms
- works well to find pathways between compounds
- Problem: What about pathways between reactions (coming from associated genes)? The atoms of which product compound should be traced?

Hub compound problem: Solution 3



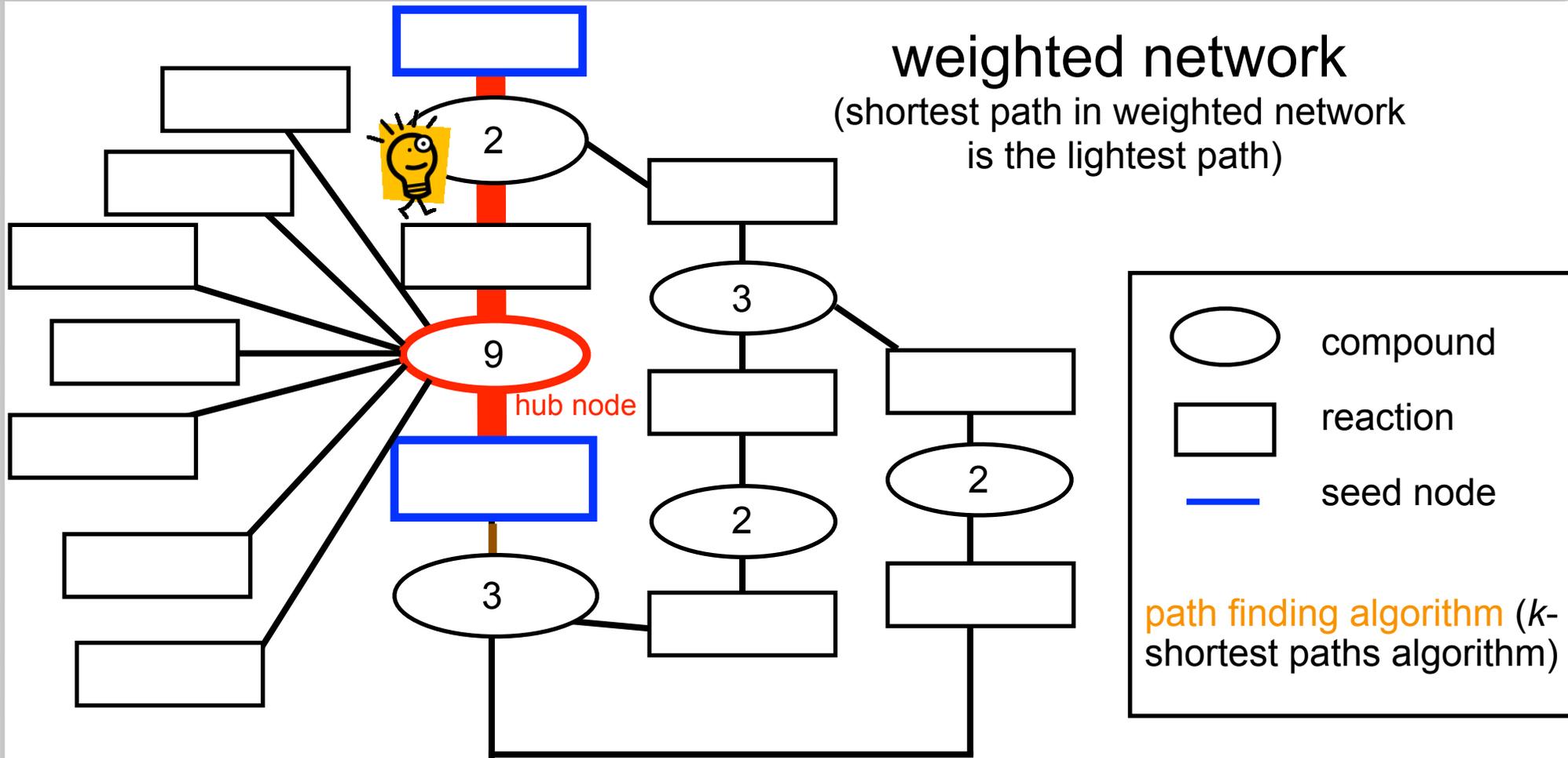
○ compound
▭ reaction
— seed node

path finding algorithm (k -shortest paths algorithm)

shortest path

weight the network (graph) to penalize hub compounds

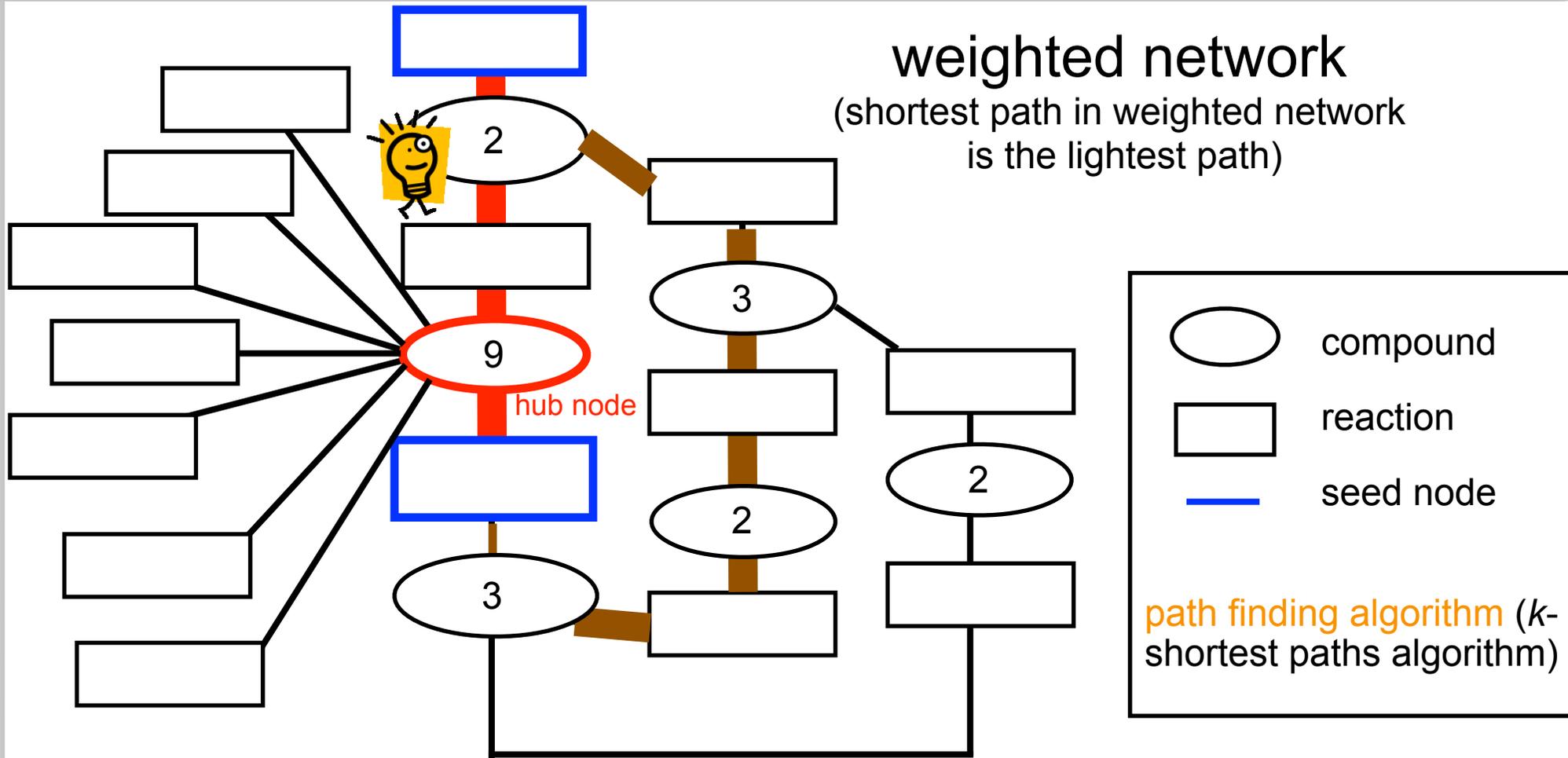
Hub compound problem: Solution 3



shortest path

weight the network (graph) to penalize hub compounds

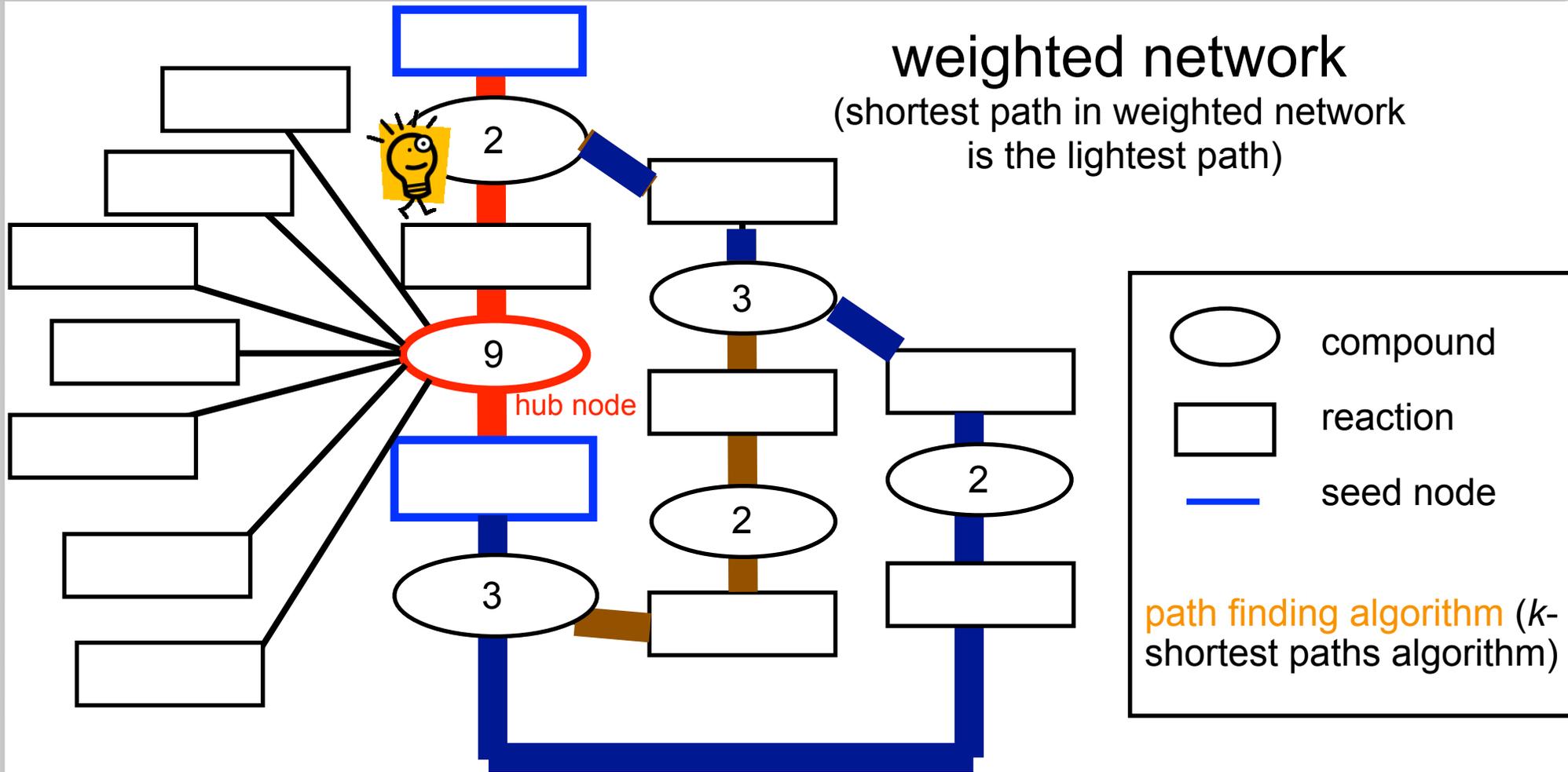
Hub compound problem: Solution 3



shortest path
lightest path

weight the network (graph) to
penalize hub compounds

Hub compound problem: Solution 3



shortest path

lightest path

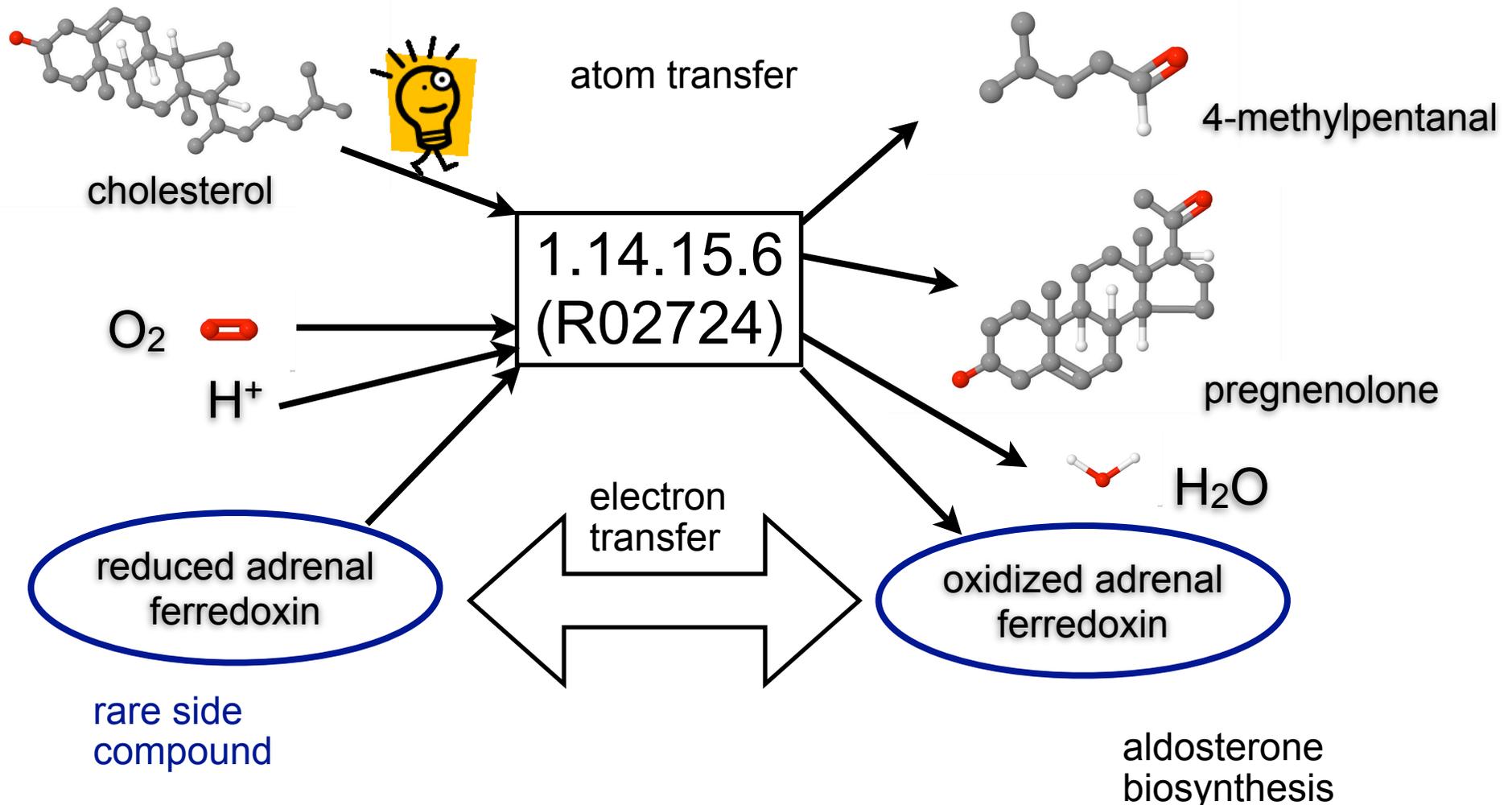
another lightest path

weight the network (graph) to penalize hub compounds

Hub compound problem: Solution 3

penalizing hub compounds with high weight works well in most cases

Problem: What about **rare side compounds**?

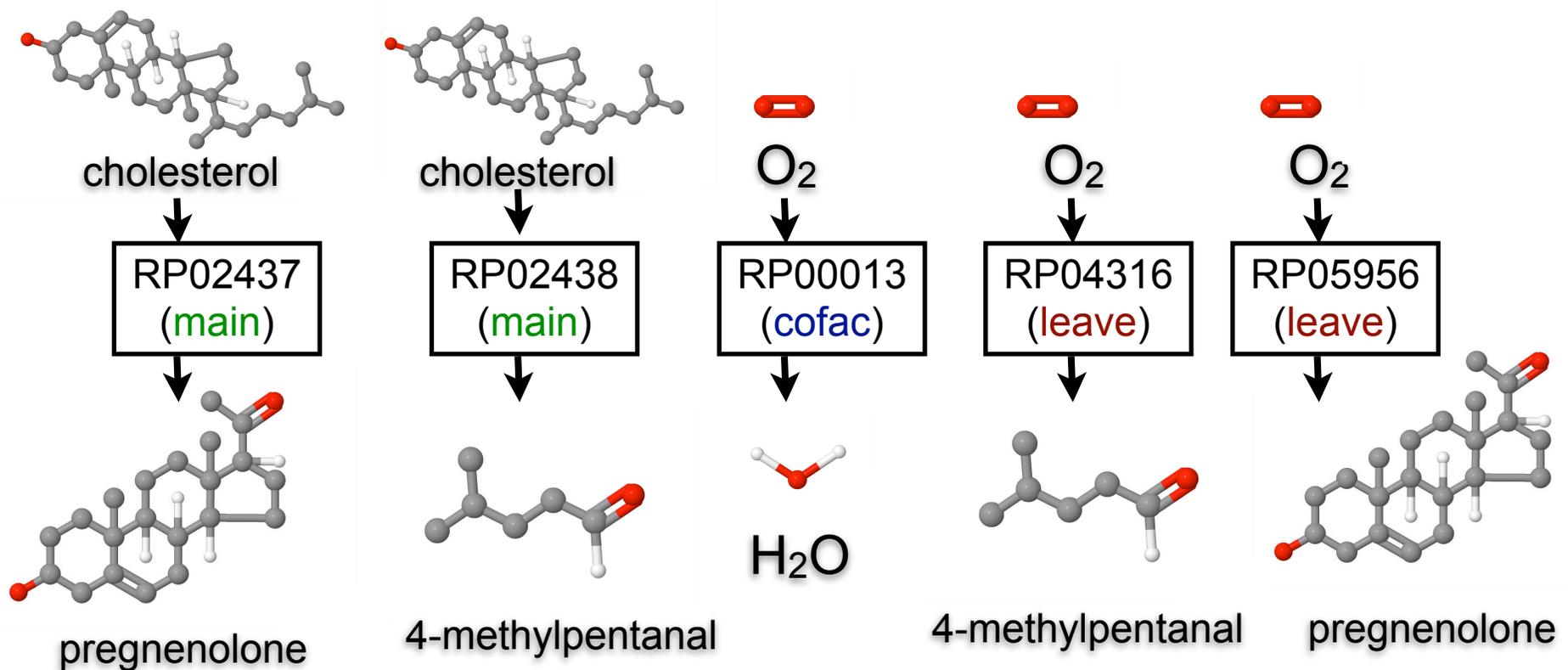


Hub compound problem: Solution 4

KEGG RPAIR database: splits reactions into reactant pairs

reactant pair: substrate and product of a reaction with high structural similarity (atom mapping)

reactant pairs have a **role** assigned such as **main**, **trans**, **cofac**, **ligase** and **leave**



Which solution works best?

Pathway prediction evaluation on 55 known pathways

Compound treatment \ Graph type	directed KEGG LIGAND	undirected KEGG RPAIR
unweighted	16%	59%
unweighted filtered (with hub compounds removed)	57%	72%
weighted	73%	83%

Conclusion: Combination of weighted network with KEGG RPAIR annotation yields highest pathway prediction accuracy

this is in agreement with work by Blum & Kohlbacher, who combined weighted network with atom mapping

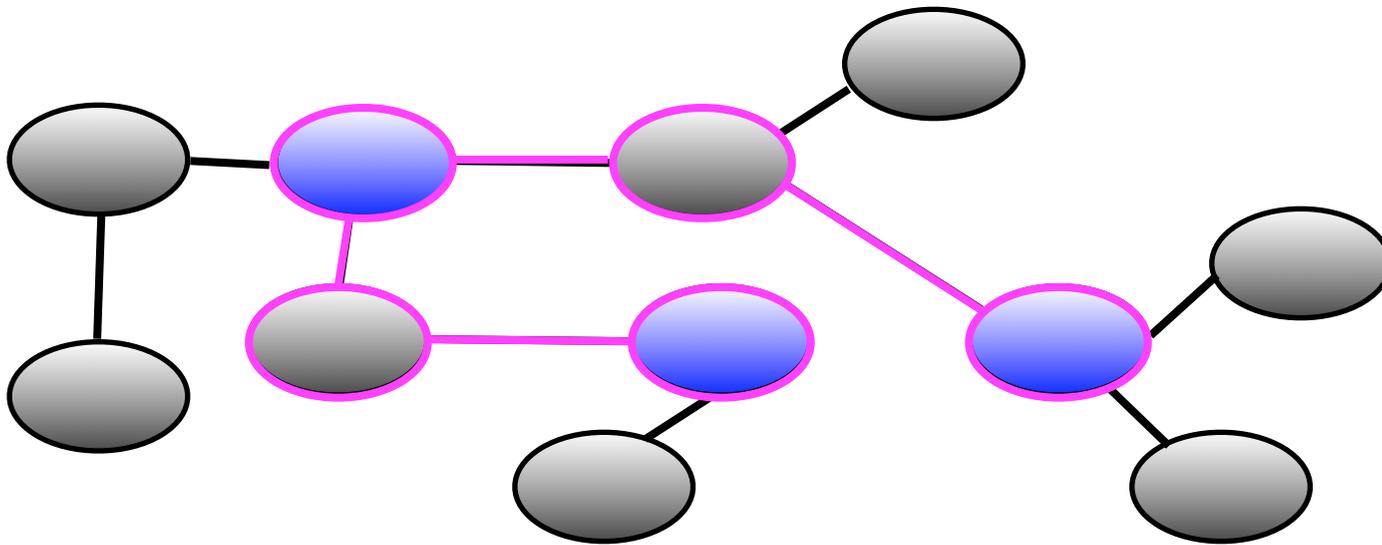
geometric accuracy in %, averaged over all predicted pathways

K. Faust, D. Croes and J. van Helden (2009). "Metabolic path finding using RPAIR annotation." *J. Mol. Biol.* 388: 390-414.

T. Blum and O. Kohlbacher (2008). "Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks." *Journal of Computational Biology*, 15: 565-576.

Subgraph extraction algorithms: Steiner tree heuristics

- Steiner tree problem: connect **seed nodes** in a graph such that the resulting **subgraph (Steiner tree)** has minimal weight



- tested three heuristics (iterative REA*, Klein-Ravi, Takahashi-Matsuyama)
- principle: calculate shortest paths repetitively and merge them

* recursive enumeration algorithm

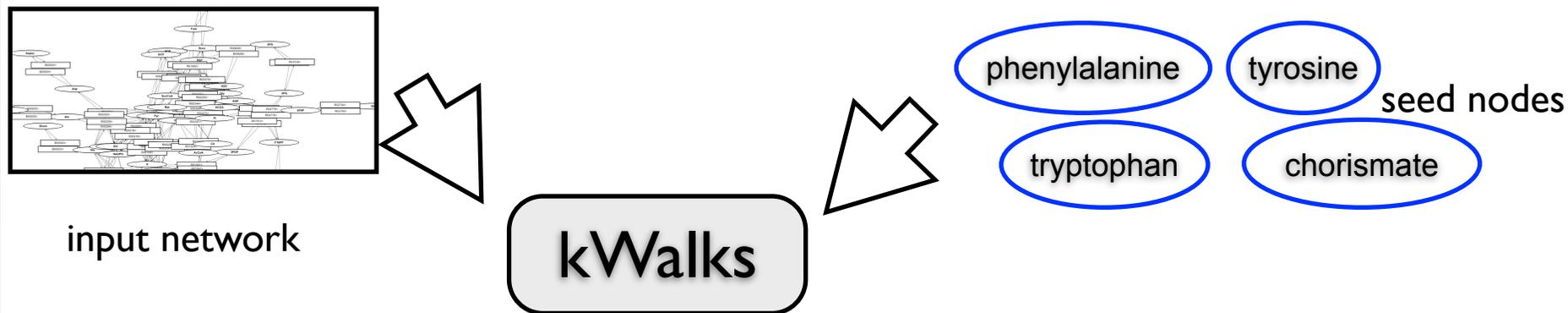
V.M. Jimenez and A. Marzal (1999). "Computing the K Shortest Paths: a New Algorithm and an Experimental Comparison." *Proc. 3rd Int. Worksh. Algorithm Engineering*, Springer Verlag.

P. Klein and R. Ravi (1995). "A nearly best-possible approximation algorithm for node-weighted steiner trees." *Journal of Algorithms*, 19:104-115.

H. Takahashi and A. Matsuyama (1980). "An approximate solution for the Steiner problem in graphs." *Math. Japonica* 24: 573-577.

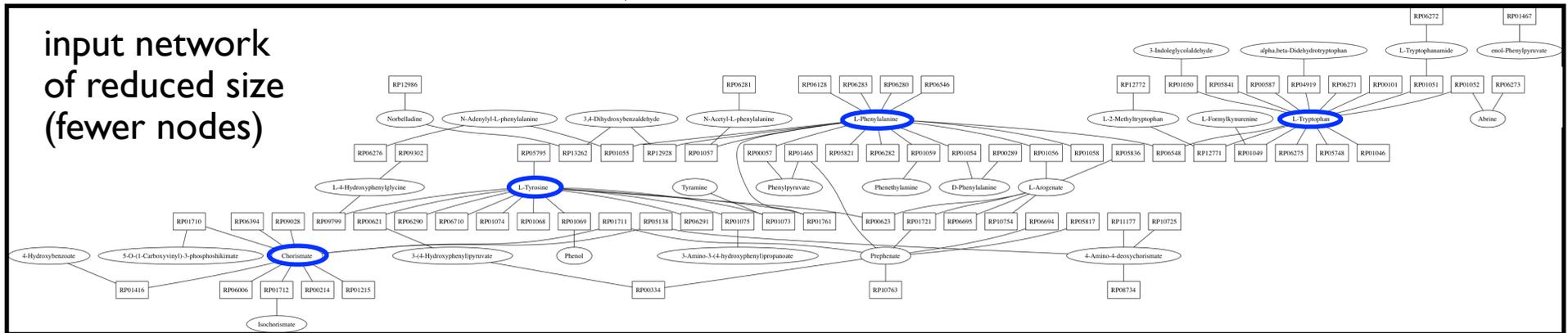
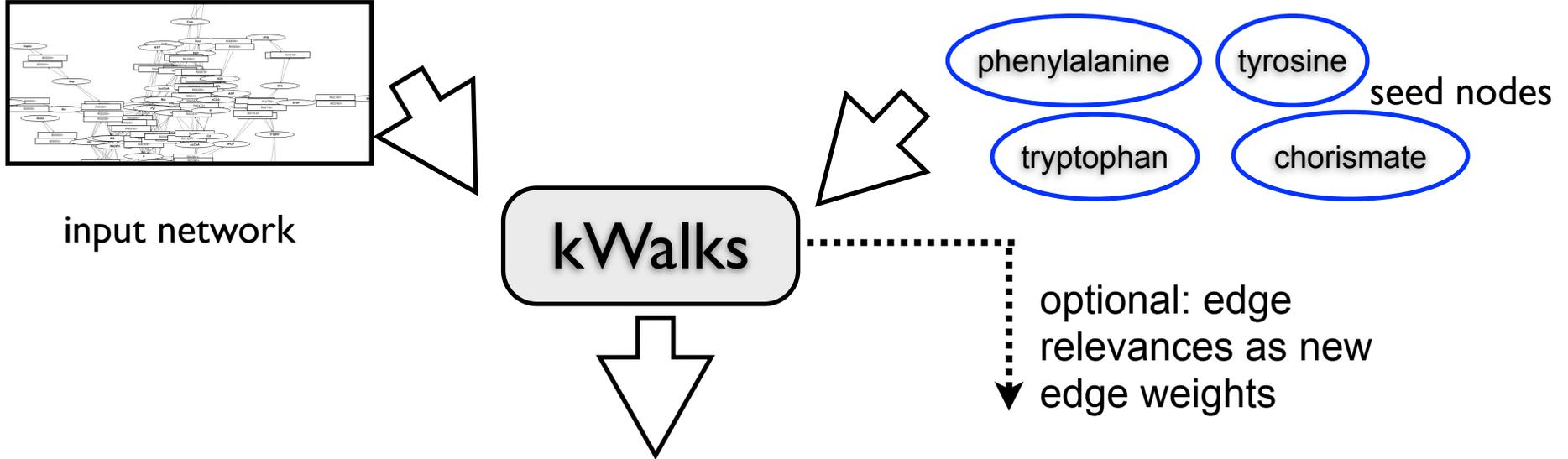
Subgraph extraction algorithms: Hybrid algorithms

- kWalks can be combined with Steiner tree heuristic



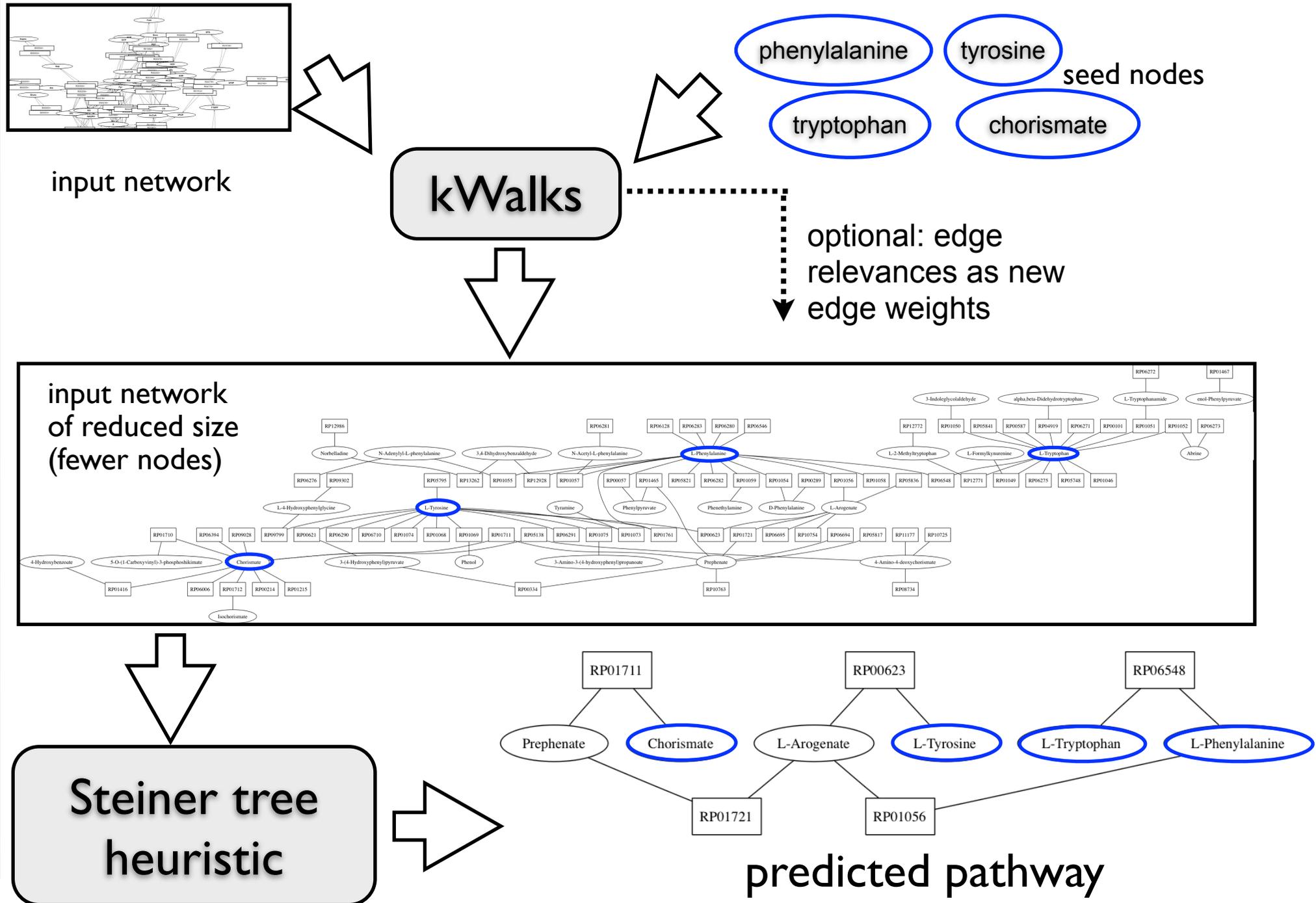
Subgraph extraction algorithms: Hybrid algorithms

- kWalks can be combined with Steiner tree heuristic



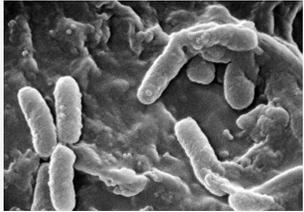
Subgraph extraction algorithms: Hybrid algorithms

- kWalks can be combined with Steiner tree heuristic



Example: *Pseudomonas aeruginosa* operon

Genes



Pseudomonas aeruginosa

Image source: Wikimedia Commons

aruCFGDB operon

PA0895 (argD)

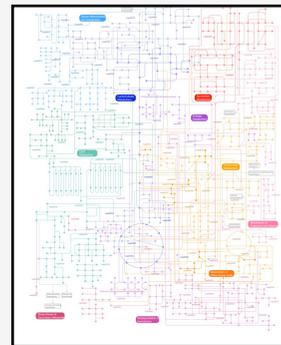
PA0896 (aruF)

PA0897 (aruG)

PA0898 (astD)

PA0899 (aruB)

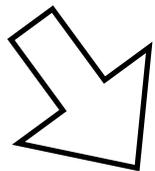
Network



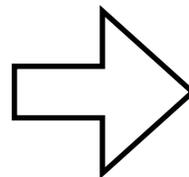
Properties:

- generic RPAIR network
- undirected, weighted
- 30,655 nodes (12,287 reactant pairs, 6,081 compounds)
- 49,148 edges
- compound nodes weighted according to their degree

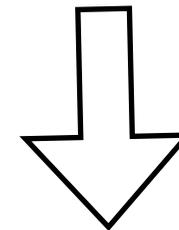
3. Example



- map genes to reactions and reactant pairs

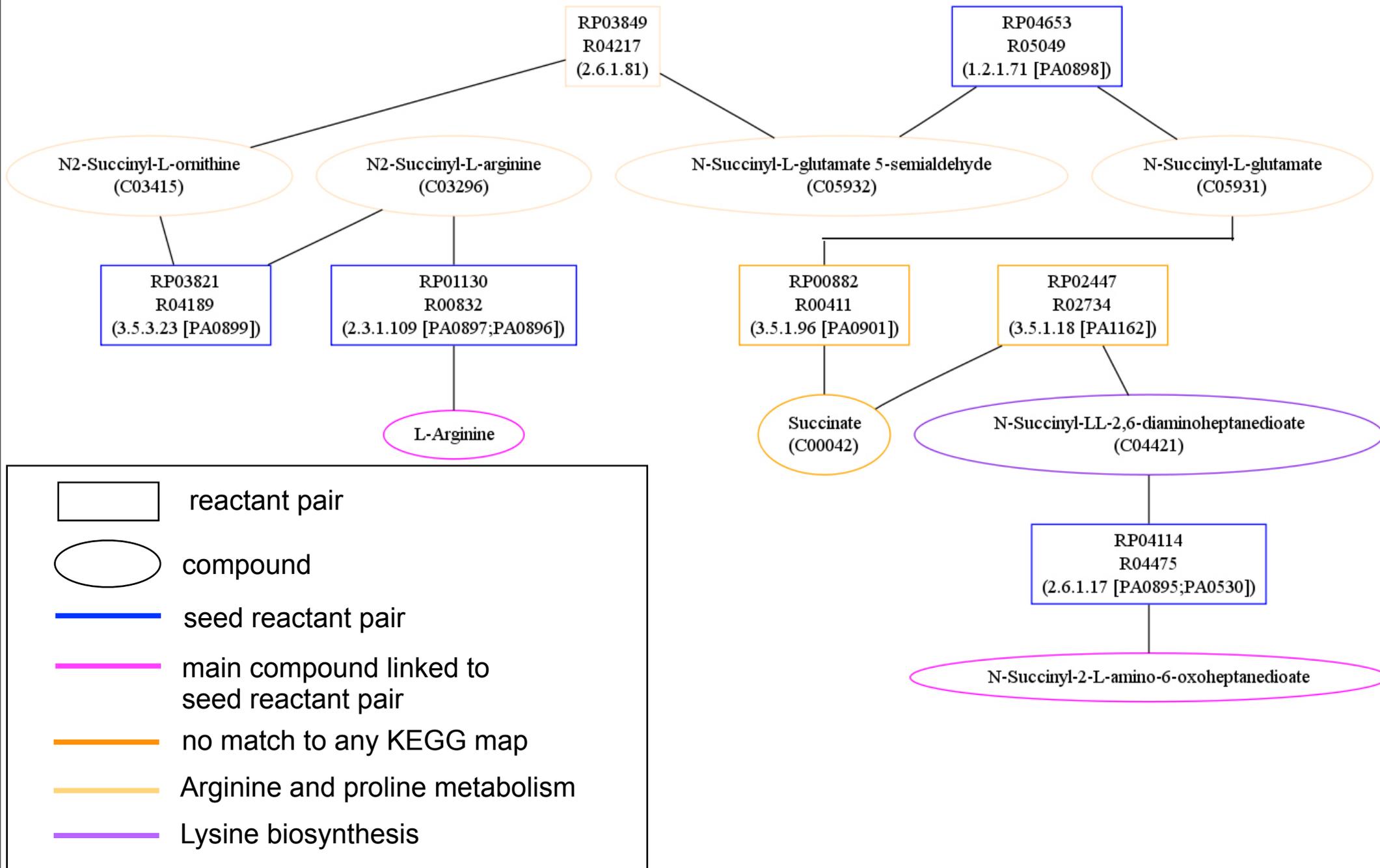


- kWalks-Takahashi-Matsuyama hybrid



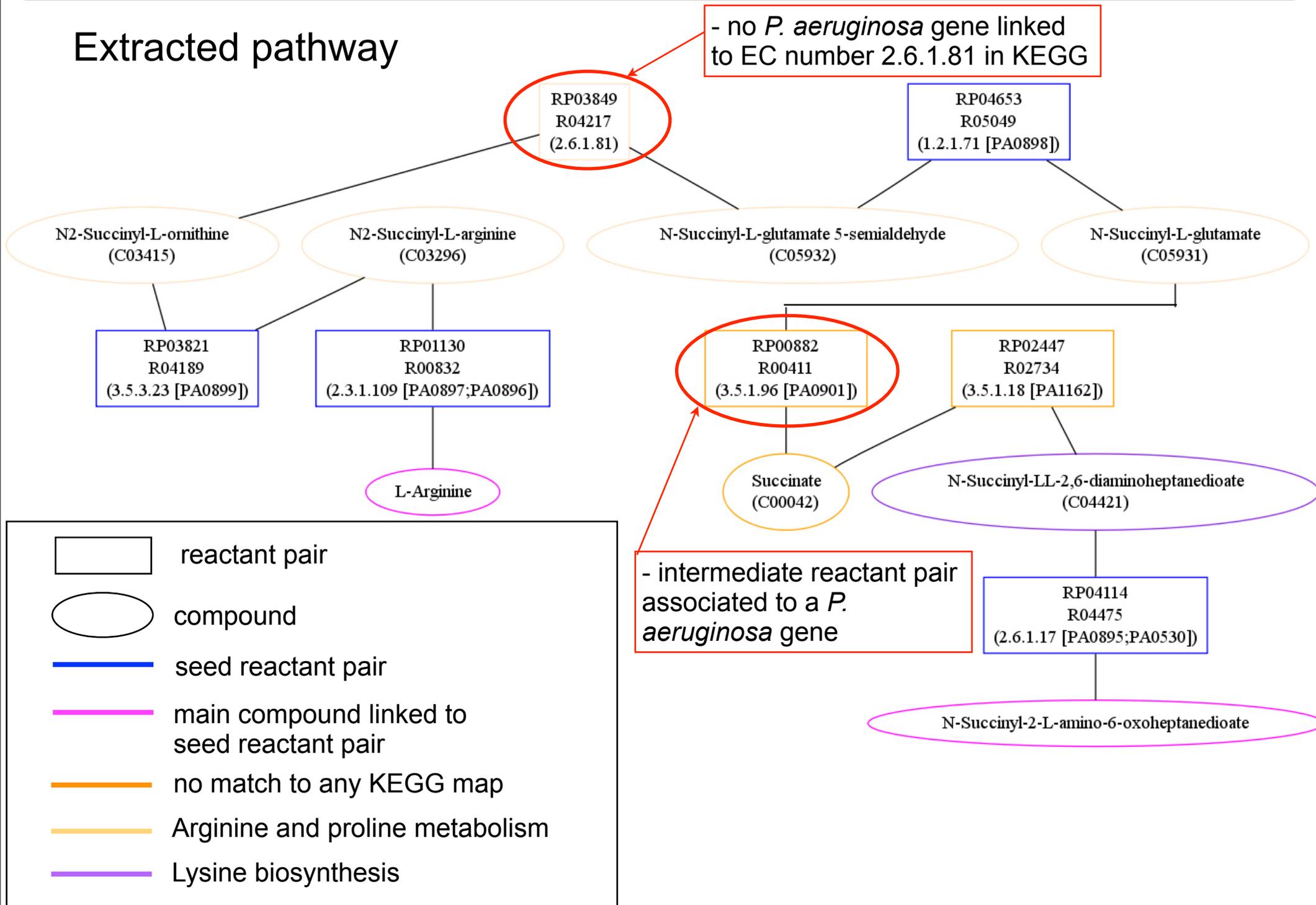
Example: *Pseudomonas aeruginosa* operon

Extracted pathway



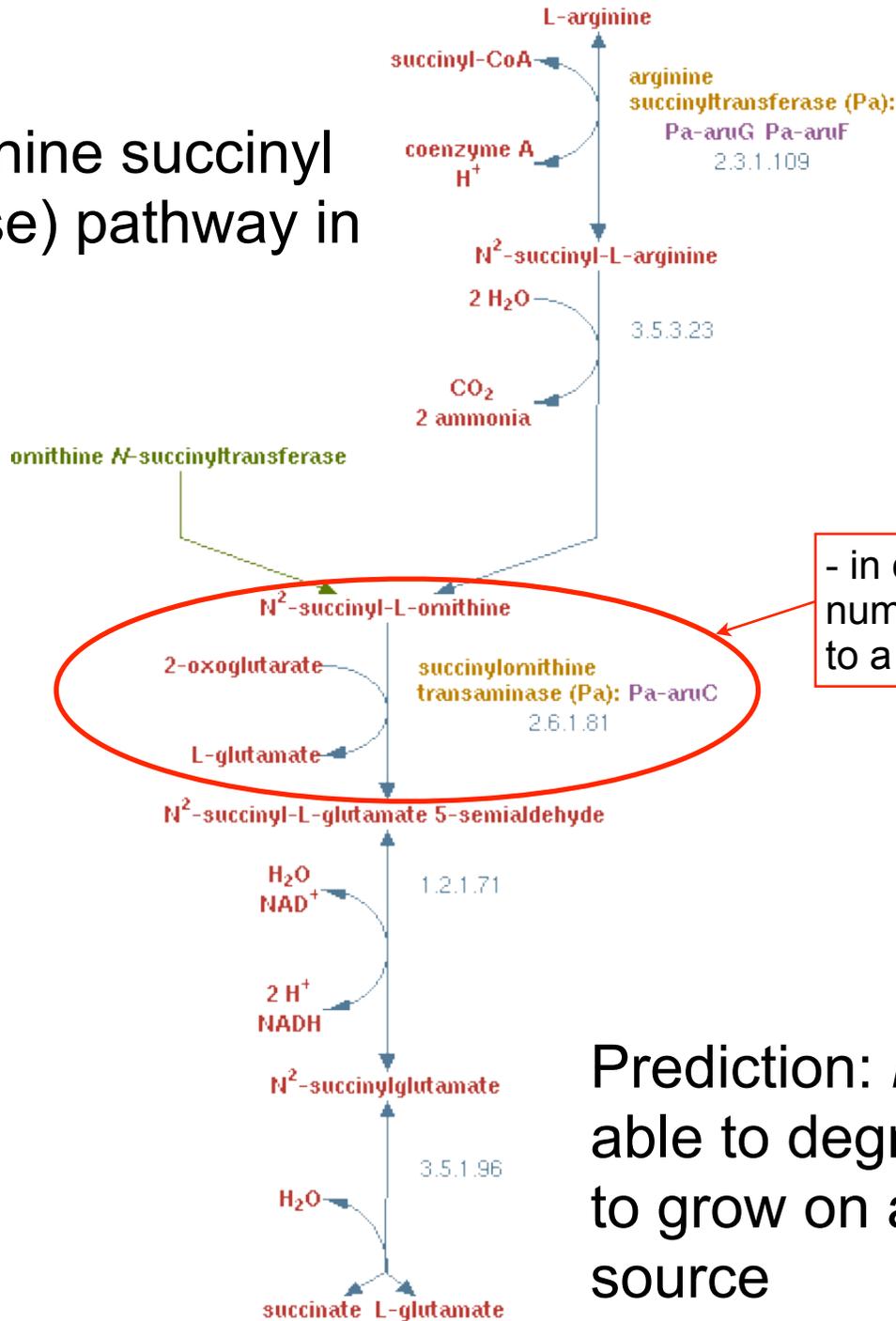
Example: *Pseudomonas aeruginosa* operon

Extracted pathway



Example: *Pseudomonas aeruginosa* operon

AST (arginine succinyl transferase) pathway in MetaCyc



- in contrast to KEGG, EC number 2.6.1.81 is linked to a *P. aeruginosa* gene

Prediction: *P. aeruginosa* should be able to degrade arginine and possibly to grow on arginine as sole carbon source

Strengths and weaknesses of pathway prediction

Strengths

-  Prediction approach can be applied to any network and handles large networks (having thousands of nodes).
-  Prediction approach only requires the network and seed nodes as input.
-  Seed nodes can be compounds or reactions/reactant pairs (EC numbers and genes).
-  Seed node sets can be treated.
-  Weights can be tuned to favor certain reactions/compounds (e.g. organism-specific reactions or reactions with high scores in a high-throughput experiment).

Weaknesses

-  Difficulty to predict pathways containing cycles or spirals (fatty acid biosynthesis).
-  Difficulty to predict pathways in highly inter-connected central metabolic network (glycolysis).
-  Difficulty to link enzymes/EC numbers to reactions.

Acknowledgement



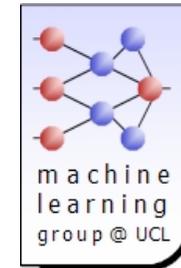
BiGRe team

IBMM

Bruno André
Patrice Godard



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



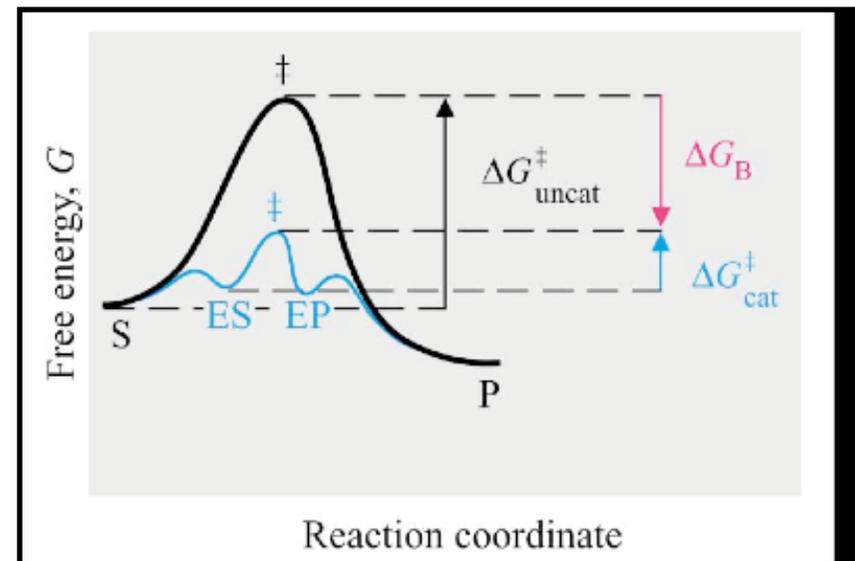
Jérôme Callut
Yves Deville
Pierre Schaus
Jean-Noël Monette

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

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:



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

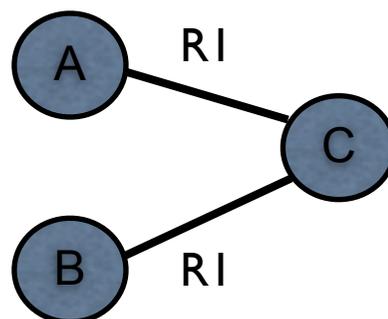
image source: <http://www.biology.buffalo.edu/courses/bio401/KiongHo/Lecture32.pdf>

Graph representation of metabolic data

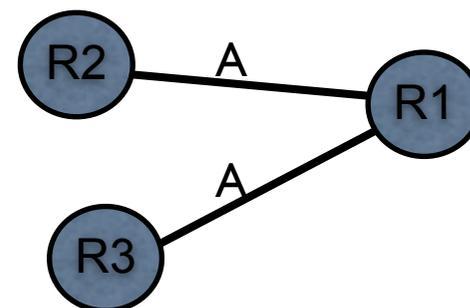
graphs with only one node set:

Why bipartite?

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



reaction R1 is represented by several edges

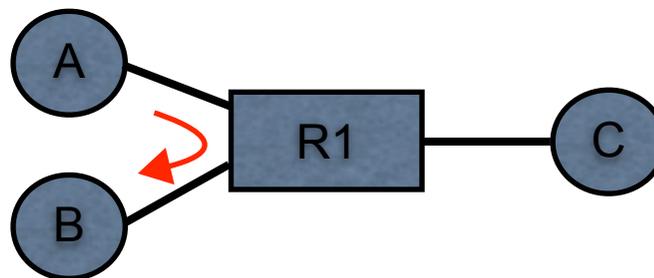


compound A is represented by several edges

Why directed?

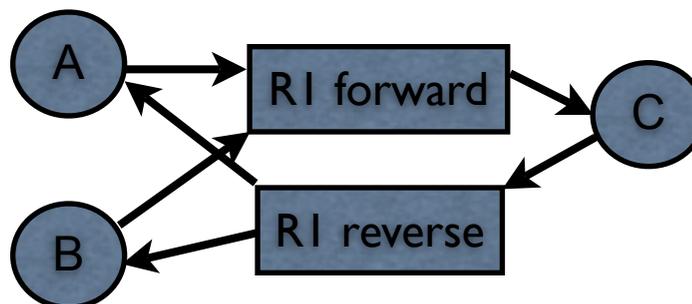
- to avoid paths going from substrate to substrate (or from product to product) of the same reaction

undirected graphs:

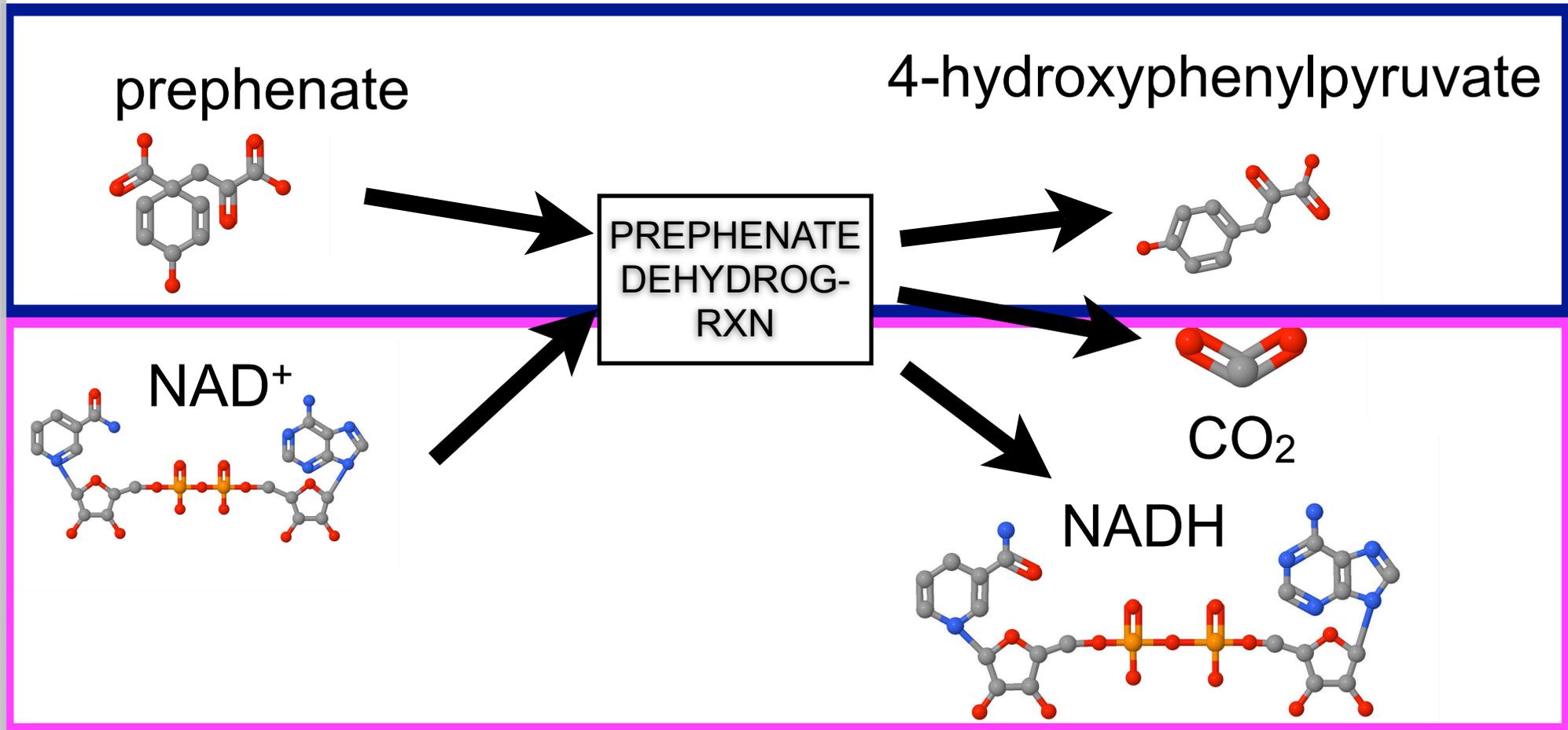


Why are direct and reverse reaction direction mutually exclusive?

- to avoid crossing the same reaction twice



Hub compound problem: Main and side compounds



main compounds: carbon atom transfer

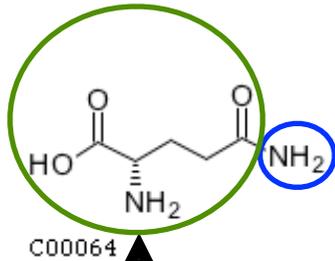
side compounds: donors/acceptors of energy, electrons or functional groups

but: distinction not always clear (e.g. glutamate)

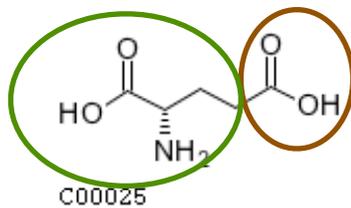
RPAIR classes

main changes
on substrate
(main)

glutamine



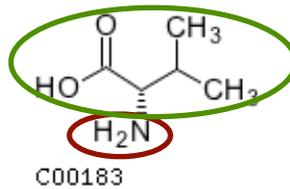
RP00024



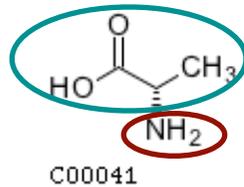
glutamate

functional groups
transferred by
transferases (trans)

L-valine

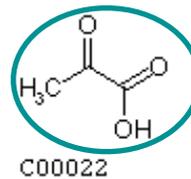


RP06488



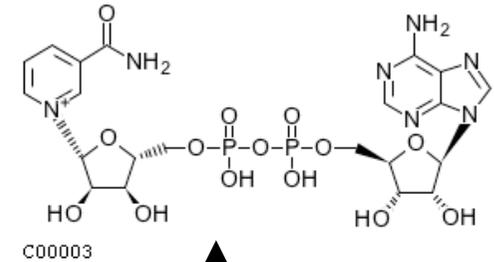
L-alanine

pyruvate

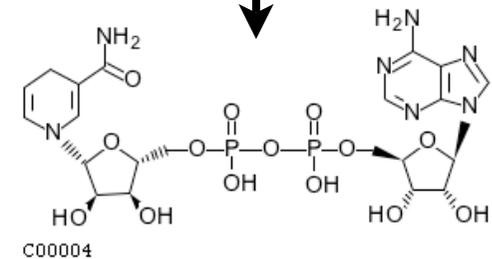


cofactor pairs in
reactions involving oxido-
reductases (cofac)

NAD+



RP00002



NADH

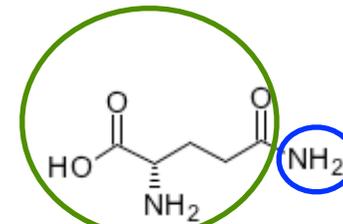
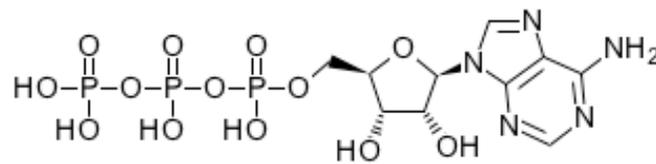
RPAIR classes

consumption of nucleoside triphosphates by ligases (ligase)

release or addition of inorganic compounds (leave)

ATP

glutamine

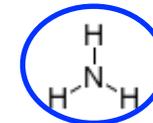
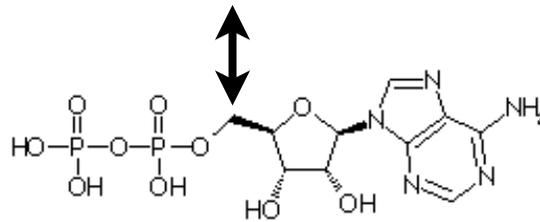


C00002

C00064

RP00003

RP05752



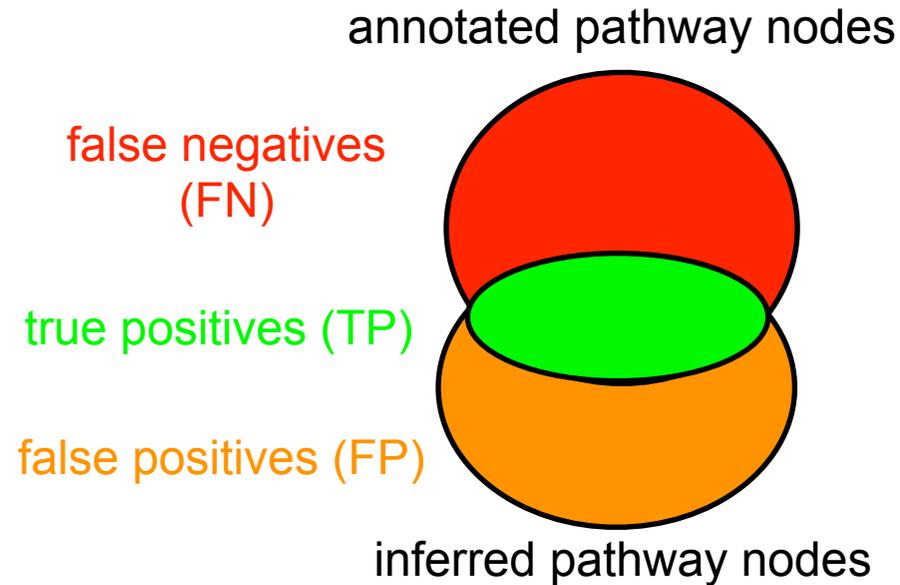
C00008

C00014

ADP

ammonia

Accuracy of pathway prediction



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

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

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

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

Multiple-end pathway prediction evaluation results

- evaluation carried out on 71 yeast-specific reference pathways in **MetaCyc** network 

Algorithm \ Weight policy	kWalks (kWalks with three iterations)	Takahashi/ Matsuyama (iterative REA)	kWalks/Takahashi- Matsuyama (iterative REA) hybrid
unweighted	62% (64%)	53% (43%)	- (55%)
weighted	60% (68%)	76% (68%)	77% (68%)

geometric accuracy in %, averaged over all predicted pathways

Seed reaction grouping problem

Genes

adhP

hisB

Enzymes

alcohol
dehydrogenase
(**broad-specificity
enzyme**)

imidazoleglycerol-
phosphate dehydratase
and histidinol-
phosphatase (**bifunctional
enzyme**)

EC numbers

1.1.1.1

4.2.1.19

3.1.3.15

Reactions

R00623

R00754

R03457

R03013

R02124

R04805

... (18)

Seed groups
(EC grouping)

R00623

R02124

R00754

... (18)

R03457

R03013

Example: *Pseudomonas aeruginosa* operon

Gene to reactant pair mapping

- N:N relationship between genes, EC numbers, reactions and reactant pairs
- seed reactant pairs can be grouped gene-wise, EC number-wise or reaction-wise

Provided identifier	Name in KEGG	Description of identifier	Associated EC numbers	Seeds used for pathway prediction	Group of seed	Identifier type
PA0899	PA0899	succinylarginine dihydrolase (EC:3.5.3.23)	3.5.3.23	[RP03821]	PA0899_group5	
PA0898	PA0898	succinylglutamic semialdehyde dehydrogenase	1.2.1.71	[RP04653]	PA0898_group4	
PA0897	PA0897	arginine/ornithine succinyltransferase AII subunit	2.3.1.109	[RP01130, RP00035]	PA0897_group3	
PA0896	PA0896	arginine/ornithine succinyltransferase AI subunit	2.3.1.109	[RP01130, RP00035]	PA0896_group2	Gene
PA0895	PA0895	bifunctional	2.6.1.17, 2.6.1.11	[RP02102, RP04114, RP00014]	PA0895_group1	Gene

- 2 genes associated to the same EC number (2 different sub-units of the same enzyme)

- bifunctional enzyme associated to 2 EC numbers

Seed enzymes come from: pae (KEGG organism abbreviation)

Seed node group treatment

- Group reactions by EC number.
- Treat each seed as a separate group.
- Keep the groups.



<http://rsat.ulb.ac.be/neat/>

P. aeruginosa example: KEGG maps overlapping with prediction

Pathways mapped to predicted subnetwork

Nodes of predicted pathway are highlighted in the KEGG map in orange (non-seed nodes) and blue (seed nodes). Organism-specific reactions are highlighted in green.

Pathway (Click to see it)	Reactions of pathway contained in the extracted subnetwork
Arginine and proline metabolism (pae00330)	[R00832 [RP01130], R00411 [RP00882], R04217 [RP03849], R04189 [RP03821], R05049 [RP04653]]
Lysine biosynthesis (pae00300)	[R04475 [RP04114], R02734 [RP02447]]

Significance of overlap between predicted subnetwork and reference pathways

ref = Reference pathway

query = Predicted pathway

R = Number of nodes in reference pathway.

Q = Number of nodes in predicted pathway.

QR = Number of nodes in the intersection of the reference and predicted node set.

QvR = Number of nodes in the union of the reference and predicted node sets.

R!Q = Number of nodes present in the reference but not in the predicted node set.

Q!R = Number of nodes present in the predicted but not in the reference node set.

jac_sim = Jaccard similarity. For 2 node sets A and B: $jac_sim = |A \cap B| / |A \cup B|$

P_val = P-value of the intersection, calculated with the hypergeometric function. $Pval = P(X \geq QR)$. The population size corresponds to the node number in the input network (16826).

E_val = E-value of the intersection. $E_val = P_val * number_of_tests$. The number of tests corresponds to the number of reference pathways in the selected metabolic database (145).

sig = Significance of the intersection. $sig = -\log_{10}(E_val)$

ref	query	R	Q	QR	QvR	R!Q	Q!R	jac_sim	P_val	E_val	sig
Arginine_and_proline_metabolism	predicted	85	15	9	91	76	6	0.09890	6.8e-18	9.86E-16	15.006
Lysine_biosynthesis	predicted	57	15	3	69	54	12	0.04348	1.6e-05	0.00232	2.635

Outlook: MICROME



- MICROME is an EU framework with the aim to establish computational and experimental pipelines for microbial pathway and network reconstruction
- contribution to computational pipeline: metabolic pathway prediction from bacterial operons and regulons