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

conditions

genes

enzyme-coding set of genes

pathway(s)

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

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**
  - b0368  aroL; shikimate kinase II [EC:2.7.1.71] [SP:AROL_ECOLI]
  - b0754  aroC; 3-deoxy-D-arabinoheptulosonate-7-phosphate synthase (DAHP synthetase, phenylalanine repressible) [EC:2.5.1.54] [SP:AROC_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]

I. Motivation - Result of pathway mapping

KEGG maps
blue: enzymes coded by query genes
grey: E. coli specific enzymes
1. Motivation - Enzymes involved in known example pathway

Superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis

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
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

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 data
metabolic graph (reaction and compound nodes)

2. Aim - Metabolic pathway inference

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

Relevant inferred metabolic pathway that may be validated experimentally

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, H₂O etc.) favor biochemically irrelevant pathways

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$_2$O etc.) favor biochemically irrelevant pathways

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, H₂O etc.) favor biochemically irrelevant pathways

- solution: weighted graph penalizing hubs

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


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$_2$O etc.) favor biochemically irrelevant pathways

- solution: weighted graph penalizing hubs

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

---


3. Methods - Definition of accuracy

reference: superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis

metabolic path finding in weighted MetaCyc graph

sensitivity $S$: $\frac{TP}{TP + FN}$
positive predictive value $PPV$: $\frac{TP}{TP + FP}$

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

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

seed reactions do not count as true positives
Pairwise $k$ shortest paths

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

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

**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

- xylulose-monophosphate cycle
- valine biosynthesis
- urea degradation
- urate degradation
- tryptophan biosynthesis
- trehalose biosynthesis
- isoleucine biosynthesis
- superpathway of sulfur amino acid biosynthesis
- superpathway of serine and glycine biosynthesis
- superpathway of ribose and deoxyribose phosphate degradation 1
- superpathway of ribose and deoxyribose phosphate degradation 2
- superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis
- superpathway of leucine, valine, and isoleucine biosynthesis
- superpathway of isoleucine and valine biosynthesis 1
- superpathway of isoleucine and valine biosynthesis 2
- superpathway of fatty acid oxidation and glyoxylate cycle 1
- superpathway of fatty acid oxidation and glyoxylate cycle 2
- glutamate fermentation I
- salvage pathways of purine nucleosides
- salvage pathways of pyrimidine ribonucleotides 1
- salvage pathways of pyrimidine ribonucleotides 2
- salvage pathways of purine ribonucleotides
- riboflavin and FMN and FAD biosynthesis
- pyridine nucleotide biosynthesis
- pyridine nucleotide cycling
- pyrimidine nucleotide biosynthesis
- polyamine biosynthesis III
- polyamine biosynthesis I
- non-oxidative branch of the pentose phosphate pathway
- methionine biosynthesis II
- methionine biosynthesis I
- serine--isocitrate lyase pathway
- serine biosynthesis
- salvage pathways of pyrimidine ribonucleotides 2
- salvage pathways of pyrimidine ribonucleotides 1
- salvage pathways of purine nucleosides
- salvage pathways of purine and pyrimidine nucleotides
- homoserine biosynthesis
- homocysteine and cysteine interconversion
- histidine biosynthesis I
- heme biosynthesis II
- glycolysis
- glutamate fermentation I--the hydroxyglutarate pathway
- glutamate degradation I
- glycine degradation
- fatty acid oxidation pathway
- de novo biosynthesis of pyrimidine ribonucleotides
- cysteine biosynthesis II
- chorismate biosynthesis
- butanedioic fermentation
- biotin pathway
- aspartate superpathway 3
- aspartate superpathway 2
- aspartate superpathway 1
- arginine degradation I
- arginine biosynthesis III
- allantoin degradation
- arginine degradation
- UMP->N-acetylglucosamine biosynthesis
- TCA cycle variant VIII
- TCA cycle --- aerobic respiration
- 4-hydroxyproline degradation
- homocysteine and cysteine interconversion
- histidine biosynthesis I
- heme biosynthesis II
- glycolysis
- glutamate fermentation I--the hydroxyglutarate pathway
- glutamate degradation I
- glycine degradation
- fatty acid oxidation pathway
- de novo biosynthesis of pyrimidine ribonucleotides
4. Evaluation of kWalks - Parameter optimization

Input graph (directed or undirected)

**Input**

seed reactions

- PSERTRANSAM-RXN→3-P-SERINE (0.043)
- 3-P-HYDROXYPYRUVATE→PSERTRANSAM-RXN (0.043)
- PGLYCDHYDROG-RXN→3-P-HYDROXYPYRUVATE (0.041)
- 2.5.1.65-RXN→CYS (0.026)
- 3-P-SERINE→2.5.1.65-RXN (0.026)

**Initial edge weights**

<table>
<thead>
<tr>
<th>edge identifier</th>
<th>edge weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSERTRANSAM-RXN→3-P-SERINE</td>
<td>0.03</td>
</tr>
<tr>
<td>2.5.1.65-RXN→CYS</td>
<td>0.01</td>
</tr>
<tr>
<td>3-P-SERINE→2.5.1.65-RXN</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**kWalks**

edge identifiers

- 3-P-SERINE
- 2.5.1.65-RXN

edge relevances

- PSERTRANSAM-RXN→3-P-SERINE: 0.043
- 3-P-HYDROXYPYRUVATE→PSERTRANSAM-RXN: 0.043
- PGLYCDHYDROG-RXN→3-P-HYDROXYPYRUVATE: 0.041
- 2.5.1.65-RXN→CYS: 0.026
- 3-P-SERINE→2.5.1.65-RXN: 0.026

iteration

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

- superpathway of phenylalanine, tyrosine and tryptophan biosynthesis
- superpathway of ribose and deoxyribose phosphate degradation
- superpathway of lysine, threonine and methionine biosynthesis
- superpathway of fatty acid oxidation and glyoxylate cycle
- non-glutamate fermentation
- salvage pathways of purine and pyrimidine nucleotides
- superpathway of isoleucine and valine biosynthesis
- oxidative branch of the pentose phosphate pathway
- salvage pathways of pyrimidine ribonucleotides
- superpathway of glycolysis and TCA variant
- superpathway of sulfur amino acid biosynthesis
- homocysteine and cysteine interconversion
- riboflavin and FMN and FAD biosynthesis
- salvage pathways of purine nucleosides
- mannosyl-N-acetylgalactosamine biosynthesis
- pyridine nucleotide biosynthesis
- monoamine biosynthesis I
- methionine biosynthesis II
- homoserine and methionine biosynthesis
- homocysteine and cysteine interconversion
- histidine biosynthesis I
- heme biosynthesis II
- glycolysis I
- glycerol degradation II
- glutamate fermentation I-the hydroxylglutarate pathway
- glutamate degradation I
- gluccoegenesis
- fatty acid oxidation pathway
- de novo biosynthesis of pyrimidine ribonucleotides
- cysteine biosynthesis II
- ornithine biosynthesis
- butanediol fermentation
- bifidum pathway
- aspartate superpathway 3
- aspartate superpathway 2
- aspartate superpathway 1
- asparagine degradation I
- arginine biosynthesis III
- allantoic degradation
- adipic acid degradation
- UDP-N-acetylglucosamine biosynthesis
- TCA cycle variation VIII
- TCA cycle — aerobic respiration
- 4-hydroxyproline degradation

number of reactions to be inferred
4. Evaluation of kWalks - Parameter optimization

Input graph (directed or undirected)

Seed reactions

Initial edge weights

kWalks

Pair-wise k shortest paths

Edge relevances

Input graph of reduced size

Edge weights

Iteration

Edge identifiers

PSERTRANSAM-RXN<→3-P-SERINE 0.03
2.5.1.65-RXN>→CYS
3-P-SERINE→2.5.1.65-RXN> 0.01

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

kWalks without iteration

kWalks with 6 iterations

kWalks (3 iterations) combined with pairwise k shortest paths

References:

- xylose-mono-phosphate cycle
- valine biosynthesis
- urea degradation
- trypophan biosynthesis
- threonine biosynthesis
- superpathway of sulfur amino acid biosynthesis
- superpathway of serine and glycine biosynthesis
- superpathway of ribose and deoxyribose phosphate degradation
- superpathway of phenylalanine, tyrosine and tryptophan biosynthesis
- superpathway of lysine, threonine and methionine biosynthesis
- superpathway of leucine, valine, and isoleucine biosynthesis
- superpathway of fatty acid oxidation and glyoxylate cycle
- salvage pathways of purine and pyrimidine nucleotides
- superpathway of isoleucine and valine biosynthesis
- oxidative branch of the pentose phosphate pathway
- salvage pathways of pyrimidine ribonucleotides
- salvage pathways of ribose and deoxyribose phosphate degradation
- riboflavin and FMN and FAD biosynthesis
- pyruvate oxidation pathway
- pyridine nucleotide biosynthesis
- non-oxidative branch of the pentose phosphate pathway
- methionine biosynthesis
- methionine biosynthesis
- mannosyl-chito-dolichol biosynthesis
- lipoygenase pathway
- leucine biosynthesis
- isoleucine degradation
- isoleucine biosynthesis
- homoserine and methionine biosynthesis
- homocysteine and cysteine interconversion
- histidine biosynthesis
- heme biosynthesis
- glycolysis
- glycerol degradation
- glutamate fermentation - the hydroxylglutarate pathway
- glutamate degradation
- glycolysis
- fatty acid oxidation pathway
- de novo biosynthesis of pyrimidine ribonucleotides
- cysteine biosynthesis
- histidine metabolism
- butanediol fermentation
- butanol pathway
- aspartate superpathway 2
- aspartate superpathway 1
- asparagine degradation
- arginine biosynthesis
- allantoin degradation
- uracil degradation
- UDP-N-acetylglactosamine biosynthesis
- polyamine biosynthesis
- TCA cycle - aerobic respiration
- 4-hydroxyproline degradation

Number of reactions to be inferred: 1 to 26

Geometric accuracy heatmap

References:

- kWalks without iteration
- kWalks with 6 iterations
- kWalks (3 iterations) combined with pairwise k shortest paths

Legend:

- 0
- 0.1
- 0.2
- 0.3
- 0.4
- 0.5
- 0.6
- 0.7
- 0.8
- 0.9
- 1
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

Jacques van Helden (supervisor)
Didier Croes
BiGRe team

IBMM
Bruno André
Patrice Godard

@MAZE
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 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:

Why directed?

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

Why weighted?

to avoid highly connected compounds

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 $\Delta G$ depends on temperature $T$ as well as on the product and educt concentration ratio and the standard free energy $\Delta G^\circ$

- these parameters are known for only a few reactions - directed metabolic graph therefore contains direct and reverse direction for each reaction

$\Delta G = \Delta G^\circ + RT \ln([\text{product}_1]...[\text{product}_n]/[\text{educt}_1]...[\text{educt}_n])$

image source: [http://www.biology.buffalo.edu/courses/bio401/KiongHo/Lecture32.pdf](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 $n_t$
  \[
  w(a) = \frac{w(n_h) + w(n_t)}{2}
  \]

**Arc weight computation kWalks**
- weight of arc $a$: inverse mean of weight of head node $n_h$ and weight of tail node $n_t$
  \[
  w(a) = \frac{2}{w(n_h) + w(n_t)}
  \]

**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

<table>
<thead>
<tr>
<th>Shortest path</th>
<th>Average sensitivity</th>
<th>Average PPV</th>
<th>Average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>31.4%</td>
<td>25.4%</td>
<td>28.4%</td>
</tr>
<tr>
<td>Filtered</td>
<td>68.0%</td>
<td>63.0%</td>
<td>65.5%</td>
</tr>
<tr>
<td>Weighted</td>
<td>88.5%</td>
<td>83.4%</td>
<td>85.9%</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Shortest path</th>
<th>Average sensitivity</th>
<th>Average PPV</th>
<th>Average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>29.6%</td>
<td>31.0%</td>
<td>29.3%</td>
</tr>
<tr>
<td>Filtered</td>
<td>63.3%</td>
<td>68.8%</td>
<td>66.6%</td>
</tr>
<tr>
<td>Weighted</td>
<td>80.7%</td>
<td>85.3%</td>
<td>83.0%</td>
</tr>
</tbody>
</table>

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

Geometric accuracy heatmap

reactions to be inferred

pathways

number of reactions to be inferred

reference pathways

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).
- A gene might be associated to a group of reactions rather than one reaction.

A gene might code for an enzyme with more than one catalytic activity. Several genes might code for enzymes with the same function (isoenzymes). A gene might be associated to a group of reactions rather than 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