BiGRe

Improved metabolic path finding using RPAIR annotation

Karoline Faust, Didier Croes, Jacques van Helden
Laboratoire de Bioinformatique des génomes et des réseaux (formerly SCMBB), Faculté des Sciences
Université Libre de Bruxelles, CP263, B-1050 Brussels, Belgium
{karoline,jvanheld}@scmbb.ulb.ac.be

Introduction

Path finding and ubiquitous compounds

The aim of path finding in metabolic networks is to obtain biochemically valid pathways, which connect a given start and end node (reaction or compound). However, compounds involved in a large number of reactions (i.e. water, ATP, NADPH etc.) hinder the detection of relevant pathways. Path finding algorithms traverse these compounds as shortcuts, inferring pathways containing co-factors or side compounds as intermediates. Different strategies have been applied to overcome this problem (i.e. exclusion of highly connected compounds [1], atom tracing [2] or rule-based methods [3]). Recently, our group introduced weighted graphs to deal with ubiquitous compounds. We evaluated path finding in the untreated, the filtered and the weighted metabolic graph (data obtained from KEGG/LIGAND [4]) and found that metabolic pathways inferred from the weighted graph reproduce the reference pathways with an average accuracy of 86%, to be compared to 66% for the filtered graph and 28% for the raw graph [5, 6]. In RPAIR [7], recently made available as part of KEGG/LIGAND, reactions have been decomposed into reactant pairs, offering another strategy to differentiate between main and side compounds.

Aim of this study

We asked whether applying metabolic path finding in a graph of sub-reactions (RPAIR) would improve the accuracy of inferred pathways. To answer this question, we compared the accuracy of pathways obtained from metabolic graphs constructed with and without RPAIR annotation. Furthermore, we evaluated the impact of additional parameters (weight, compound filtering, graph structure) on path finding accuracy.

Methods

Graph construction

The following bipartite graphs have been constructed from KEGG/LIGAND (Release 41.0):

A) Reaction graph

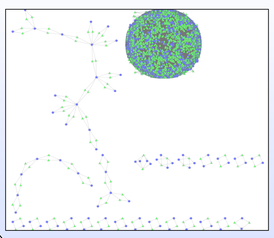
The reaction graph has been built from all compounds and reactions listed in KEGG/LIGAND (excluding glycans).

B) Sub-reaction graph

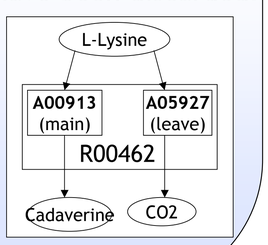
The sub-reaction graph is constructed from all reactant pairs present in RPAIR as sub-reactions and from all compounds that are educt (substrate) or product of a sub-reaction.

C) Reaction-specific sub-reaction graph

This graph has been constructed the same way as the reaction graph, but with each reaction divided in its sub-reactions. Reactions not listed in RPAIR have been discarded.



Left: The reaction graph has 18,030 nodes and 53,572 arcs and consists of 25 weakly connected components.
Right: Example for the division of a reaction into two reactant pairs (sub-reactions).



Path finding

REA has been used as k shortest paths algorithm [8]. The inferred pathway is the union of all paths of first rank obtained for a given start and end reaction.

Reference pathway set

E. coli metabolic pathways obtained from aMAZE [9] have been filtered (no cycles, at least 5 nodes, reactions present in RPAIR database), resulting in a test set of 32 pathways.

Parameter

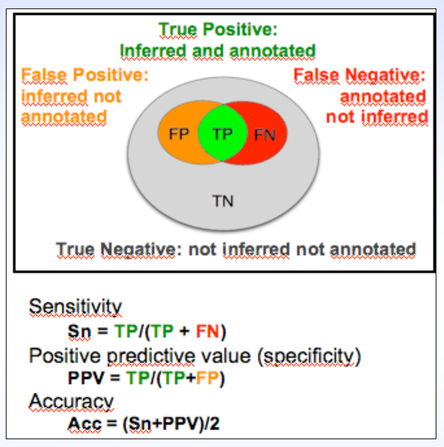
The following parameter values have been tested:

Parameter	Values
Graph type	A, B, C
Allowed RPAIR type*	main, main and trans, all
Reaction weights	according to RPAIR type, none
Compound weights	degree of compound, none
Direction	directed, undirected
Filtering	removal of 36 most connected compounds, no filtering

*There are five RPAIR types: main, trans, cofac, ligase and leave.

Evaluation

Calculation of path finding accuracy



Evaluation procedure

For each pathway from the reference set:

- identify start and end reaction
- do path finding given start and end node
- compare inferred pathway to reference pathway and calculate accuracy

Results

Optimal parameter combinations

We evaluated each of the 104 possible parameter combinations. Below, we show four from the top-ranking combinations (accuracies averaged over all pathways):

Graph type	RPAIR types	Reaction weights	Compound weights	Directed graph	Filtered graph	Geometric accuracy
B	main/trans	RPAIR	degree	true	false	0.93
B	main/trans	none	degree	true	false	0.93
B	all	RPAIR	degree	true	false	0.93
B	all	RPAIR	degree	false	false	0.93

For comparison, the accuracies obtained with combinations corresponding to our previous study are listed below (note that a more recent version of KEGG/LIGAND than in [5, 6] has been used):

Graph type	RPAIR types	Reaction weights	Compound weights	Directed graph	Filtered graph	Geometric accuracy
A	none	none	none	true	false	0.19
A	none	none	none	true	true	0.69
A	none	none	degree	true	false	0.81

Study cases

Below, two example pathways illustrate how including RPAIR annotation improves path finding accuracy. The graphs tested differ only in one parameter (graph type).

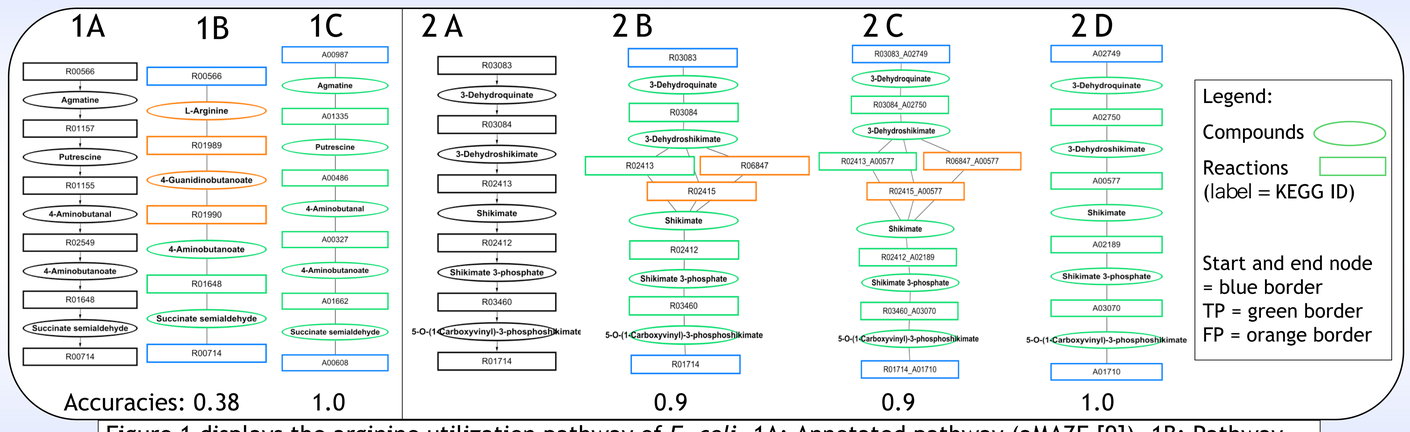


Figure 1 displays the arginine utilization pathway of *E. coli*. 1A: Annotated pathway (aMAZE [9]). 1B: Pathway obtained from weighted reaction graph. 1C: Pathway inferred from sub-reaction graph

Figure 2 shows the chorismate biosynthesis pathway of *E. coli*. 2A: Annotated pathway (aMAZE [9]). 2B: Pathway inferred from weighted reaction graph. 2C: Pathway inferred from reaction-specific sub-reaction graph. 2D: Pathway obtained from sub-reaction graph.

Conclusion

We evaluated the impact of a number of parameters (graph type, allowed RPAIR types, weighting scheme, filtering of ubiquitous compounds) on the accuracy of pathways inferred from the KEGG graph. We observed that highest average accuracy (93%) for the tested pathway set is reached when using a metabolic graph constructed with RPAIR and a weighting scheme penalizing highly connected compounds. Thus, our study confirms that the biochemical knowledge represented by the RPAIR database improves the accuracy of metabolic pathway predictions. In the near future, we will benefit from these methodological improvements to interpret gene co-regulation networks with the help of path finding.

References

[1] J. van Helden, L. Wernisch, D. Gilbert and S. Wodak (2002) **Graph-based analysis of metabolic networks**. Bioinformatics and Genome Analysis, Vol. 38, Springer-Verlag
[2] M. Arita (2004) **The metabolic world of Escherichia coli is not small**. PNAS Vol. 101(6), 1543-1547
[3] B.K. Hou, L. Ellis and L. Wackett (2004) **Encoding microbial metabolic logic: predicting biodegradation**. J. Ind. Microbiol. Biotechnol. Vol. 31, 261-272
[4] M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya (2002) **The KEGG databases at GenomeNet.** Nucleic Acids Research Vol. 30 (1), 42-46
[5] D. Croes, F. Couche, S. Wodak and J. van Helden (2005) **Metabolic PathFinding: inferring relevant pathways in biochemical networks**. Nucleic Acids Research, Vol. 33, W326-W330
[6] D. Croes, F. Couche, S. Wodak and J. van Helden (2006) **Inferring Meaningful Pathways in Weighted Metabolic Networks**. J. Mol. Biol., 356, 222-236
[7] M. Kotera et al., (2004) **RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions.** Genome Informatics 15
[8] 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
[9] C. Lemer, H. Anerhour, J.M. Maniraja, O. Sand, J. Richelle and S. Wodak (2004) **The aMAZE database goes public.** ECCB

Acknowledgements

Karoline Faust is supported by: The Actions de Recherches Concertées de la Communauté Française de Belgique (ARC grant number 04/09-307).

We would like to thank our collaborators from the BioMaze/TransMaze projects: the INGI team (UCL) for giving us access to their graph analysis tools and the former aMAZE team, who implemented software to represent, store and analyze metabolic pathways. Thanks as well to the BiGRe team for helpful discussions.