Predicting metabolic pathways from bacterial operons and regulons

Karoline Faust[1], Didier Croes[2] and Jacques van Helden[2]

1) Bioinformatics and Systems Biology (BSB), (http://systemsbiology.vub.ac.be), VIB/Vrije Universiteit Brussel, B-1050, Brussels, Belgium
2) Network and Genome Bioinformatics (BiGRe), (http://www.bigre.ulb.ac.be/), Université Libre de Bruxelles, B-1050, Brussels, Belgium

Introduction

A number of experimental and bioinformatics analyses results in sets of co-regulated, co-expressed or co-occurring enzyme-coding genes. Our aim is the prediction of metabolic pathways from these enzyme-coding genes, which are assumed to be functionally related. In contrast to pathway matching approaches, pathway discovery can detect variants or combinations of known metabolic pathways.
In addition, it can be applied to organisms whose metabolism is unknown, but for which sets of functionally related, annotated genes are available.

Methods

The idea of pathway discovery is the following: Given a set of seed reactions and a generic or organism-specific metabolic network, a sub-network is extracted by connecting the seeds in the input network. This sub-network represents the predicted metabolic pathway [1].
We have recently evaluated a number of sub-network extraction algorithms on known pathways [2] and found that the combination of a random walk-based approach [3] with a Steiner tree heuristic [4] yields reasonable prediction accuracies.
A major problem of pathway prediction are hub compounds such as ATP or H_2O, which are involved in hundreds of reactions. Naïve graph algorithms will traverse these hub compounds preferentially, thus predicting biochemically unrealistic pathways. We deal with the hub compound problem in two ways: (1) by weighting the metabolic network [5] and (2) by employing the KEGG RPAIR database [6], which provides reaction-specific main/side compound annotations. We found that the accuracy achieved with a combination of these approaches is superior to that of each approach alone [7].
When predicting pathways from enzyme-coding genes, the input genes have to be associated to reactions. This is not an easy task, as genes, EC numbers and reactions are linked by a many-to-many relationship. In addition, it is not clear whether input reactions obtained from genes should be grouped gene-wise, EC number-wise or reaction-wise. We prefer the EC number-wise grouping of reactions, because multi-functional enzymes may contribute several reactions to a pathway, whereas reaction-wise grouping introduces too many irrelevant reactions in case of imprecise gene-reaction mappings.
We developed a pathway extraction tool, which accepts a set of enzyme-coding genes, links them to reactions and predicts a pathway from these.

Study case *Pseudomonas aeruginosa PAO1*

aruCFGDB **operon** containing genes:
PA0895, PA0896, PA0897, PA0898, PA0899

1. The genes are given as input to the pathway extraction tool, which is a part of the Network Analysis Tool suite (NeAT) [8].

2. Result of gene to reaction mapping.

The table below lists information on each of the given seed node identifiers.
Green means that the given seed node identifier has been found in the database.
Orange means that there are several matches for the given seed node identifier and that you have to choose the appropriate one. For instance, if you give "lysine" as seed node identifier, should it be L-Lysine or D-Lysine?
Red means that the given seed node identifier could not be found in the database and will be neglected.

Reactions are mapped to their main reactant pairs.

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	Gene
PA0898	PA0898	succinylglutamic semialdehyde dehydrogenase	1.2.1.71	[RP04653]	PA0898_group4	Gene
PA0897	PA0897	arginine/ornithine succinyltransferase AII subunit	2.3.1.109	[RP01130, RP00035]	PA0897_group3	Gene
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

Seed enzymes come from: pae (KEGG organism abbreviation)

Seed node group treatment [?]

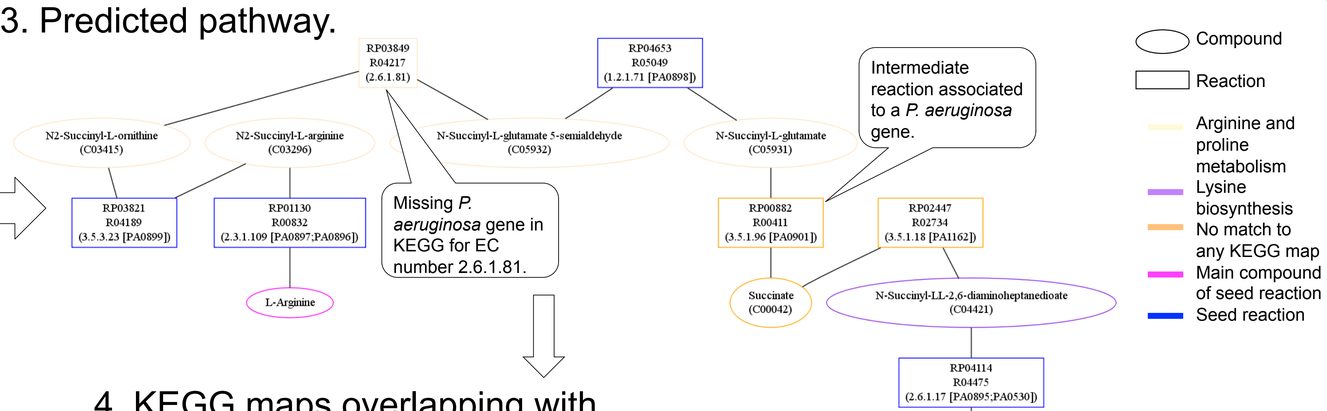
- (•) Group reactions by EC number.
- ( ) Treat each seed as a separate group.
- ( ) Keep the groups.

Input reactions can be grouped gene-wise ("Keep the groups"), EC number-wise or reaction-wise ("Treat each seed as a separate group").

PA0897 and PA0896 are associated to the same EC number, because they code for sub-units of the same enzyme.

PA0895 codes for a bifunctional enzyme associated to two EC numbers.

3. Predicted pathway.



4. KEGG maps overlapping with predicted pathway.

Pathways mapped to predicted subnetwork

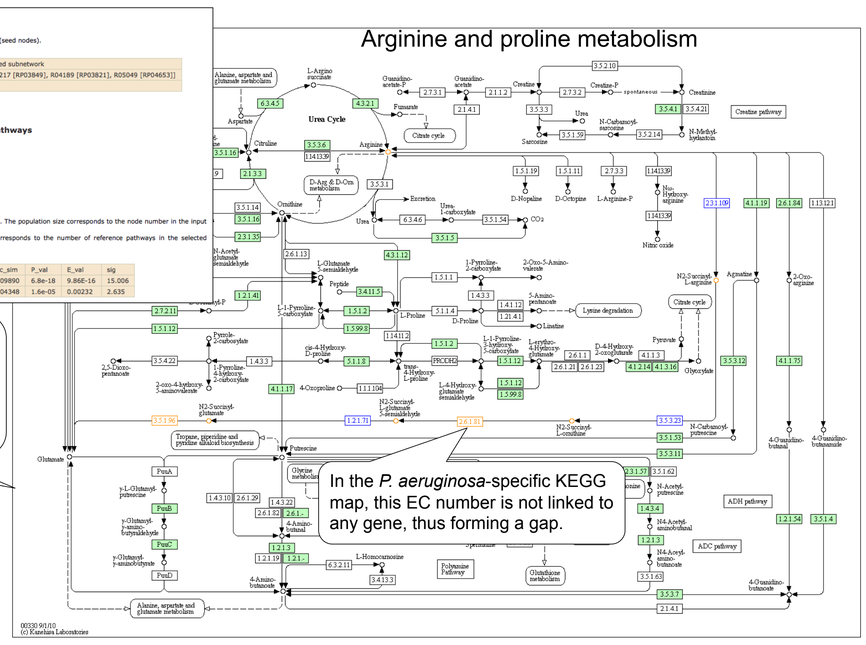
Nodes of predicted pathway are highlighted in the KEGG map in orange (non-seed nodes) and blue (seed nodes).
Organism-specific reactions are highlighted in green.

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.
RIQ = Number of nodes present in the reference but not in the predicted node set.
QIR = 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 intersection B| / |A union B|)
P_val = P-value of the intersection, calculated with he hypergeometric function. Pval = P(X >= 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 = -log10(E_val)

ref	query	R	Q	QR	QvR	RIQ	QIR	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



The predicted pathway is highlighted in orange (intermediate nodes) and blue (seed nodes) in the *P. aeruginosa* -specific KEGG map of arginine and proline metabolism. EC numbers associated to *P. aeruginosa* genes are colored in green.

In the *P. aeruginosa*-specific KEGG map, this EC number is not linked to any gene, thus forming a gap.

The predicted pathway reproduces correctly the arginine succinyl-transferase (AST) pathway.

Conclusions & Perspectives

Our methodology predicts pathways from functionally related, enzyme-coding genes and can thus propose pathways for organisms with unknown metabolism but known operons and regulons. In case of organisms where operons and regulons are not yet annotated, the idea is to predict them by combining pattern discovery in bacterial promoters with phylogenetic footprinting.
This approach will be applied within the MICROME project (http://www.microme.eu), which will establish pipelines involving both computational and experimental approaches in order to assemble metabolic pathways and to reconstruct metabolic networks in bacteria.

References

1. van Helden, J., Wernisch, L., Gilbert, D. & Wodak, S. (2002). Graph-based analysis of metabolic networks. Ernst Schering Res. Found. Workshop 38, 245-274.
2. Faust K., Dupont, P., Callut, J. & van Helden, J. (2010). Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics 26, 1211-1218.
3. Dupont, P., Callut, J., Dooms, G., Monette, J.-N., & Deville, Y. (2006). Relevant subgraph extraction from random walks in a graph. Research Report UCL/FSA/INGI RR 2006-07.
4. Takahashi, H. and Matsuyama, A. (1980). An approximate solution for the Steiner problem in graphs. Math. Japonica 24, 573–577.
5. Croes, D., Couche, F., Wodak, S. & van Helden, J. (2005). Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Research 33, W326-W330.
6. Kotera, M., Hattori, M., Oh, M.-A., Yamamoto, R., Komeno, T., Yabuzaki, J., Tonomura, K., Goto, S. & Kanehisa, M. (2004). RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. Genome Informatics 15, P062.
7. Faust, K., Croes D. & van Helden, J. (2009). Metabolic Pathfinding Using RPAIR Annotation. Journal of Molecular Biology 388, 390-414.
8. Brohée S., Faust, K., Lima-Mendez G., Sand, O., Janky, R., Vanderstocken G., Deville, Y. & van Helden, J. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. Nucleic Acids Research 36, W444-W451.

Availability

NeAT is available at: http://rsat.ulb.ac.be/neat/

Acknowledgements

KF's work was supported by Actions de Recherches Concertés de la Communauté Française de Belgique (ARC grant number 04/09- 307). DC is supported by the MICROME project (EU 7th Framework Program, Grant Agreement Number 222886-2). The BiGRe laboratory is a member of the BioSapiens Network of Excellence funded under the sixth Framework program of the European Communities (LSHG-CT-2003-503265).This work was partly funded by the Belgian Federal Science Policy Office (IAP P6/25, BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011).