CONNECTIVITY ANALYSIS OF METABOLITES IN SYNTHETIC METABOLIC PATHWAYS

Jurijs Meitalovs, Egils Stalidzans
Latvia University of Agriculture
jurijs.meitalovs@gmail.com, egils.stalidzans@gmail.com

Abstract. The size and number of available genome scale metabolic models and reconstructions is growing rapidly. Various computational methods and tools are developed in the last decades for metabolic network construction, analysis and optimization. A software tool *SpaceAnalyzer* (*SAnalyzer*) is developed in Matlab environment for automatic generation of biochemical pathways to new metabolites. The computational experiments were performed to build automatically the connectivity tree for 2,3-Butanediol, *n*-Butanol, caprolactam, lactose, penicillin and cellulose as substrates. The software tool is freely available online at: http://www.biosystems.lv/sanalyzer.

Keywords: synthetic biology, solution space, metabolic pathway construction, metabolism.

Introduction

Metabolic pathways consist of sequences of biochemical reactions acting in a microorganism. Biotechnological pathway of reactions converts the available substrate via a set of reactions into a more valuable biotechnological product. A reaction can happen if it is enabled by genetically encoded enzyme that can be inserted into an organism.

Production of different substances using microorganisms is becoming more popular in last decades. It is possible to use genetically modified organisms to produce valuable substances on an industrial scale with a lower cost price. Metabolic Engineering (ME) usually is applied for optimizing regulatory and genetic processes in cells to increase the yield of the biotechnological process [1-5]. ME has developed towards detailed metabolic analyses to identify targets for genetic manipulation and modification to improve or design cells thus overlapping with the field of synthetic biology [6; 7] serving biofuels, biomedicals, environmental and other industries.

The reconstruction of the networks of biochemical pathways in organisms based on the genome sequencing and annotation results [8-10] can be used for development and analysis of ME relevant modifications.The reconstructions and reconstruction based models of the microorganisms can be represented in a form of a graph [11-13] and can be used in design of new biological pathways. Reconstructed networks and their models are available online in the biological databases, e.g., BIGG [14], EcoCyc [15], BioCyc [16], TheSEED [17] and others. New models can be combined from the existing ones of the same organism [18; 19].

Many successful reconstructions and models are developed [3; 5; 20-22] but just very few methodologies can effectively aid in the rational design of microbial strains. Most of the proposed analysis approaches are based on the experience of biologists without systematic analysis of all possible solutions due to the high number of alternatives. Some attempts to construct metabolic pathways in the automated way are presented [23; 24].

Automatic pathway construction tool *Sanlyzer* [25] is demonstrated. This tool creates a connection matrix of metabolites based on scanning of the complete space of the possible pathways enabled by reactions registered in databases. The dynamics and numerical characteristics of automatic generation of the pathway tree are analyzed for different substrates using *SAnlyzer.*

SAnalyzer tool

*SAnalyzer*tool is a set of algorithms for metabolic pathways construction [25]. *SAnalyzer* reads all necessary data from KEGG database [26] REST web service and processes them in the real time. It uses Matlab *geturl* function to read the data using KEGG API and parses it. KEGG API allows searching biological data about biochemical networks. KEGG database keeps data about almost 17000 metabolites, molecules and other chemical substances and more than 9000 reaction entries taken from the metabolic pathway maps. Each entry has KEGGID identifier, e.g., “C00243” for lactose compound. All reaction identifiers are starting with “R”. KEGG database is provided only for academic purposes and this service cannot be used for data downloads.

Algorithms are realized in *M* scripts that can be places on the local PC and launched from the Matlab command windows. The construction algorithm of the possible metabolic pathways construction is independent software that can return nodes and edges of the constructed graph. The tree construction proceeds recursively, beginning from the starting metabolite, i.e., initial substrate of the pathway of interest given by the industrialist, as a root of a tree. We can call it "super graph" which contains all possible pathways from the input substrate for fixed levels count. Construction is performed based on the graph theory methods.

For each metabolite all reactions where it is involved are found in the data base. For each reaction the main reaction pairs in KEGG data base are chosen where the active metabolite is involved. From these reaction pairs the list with metabolites for the next construction level is created. If the metabolite is already present in the graph, it is not used in the following reactions. That prevents appearance of loops and reduces the number of iterations because only unique metabolites are presented in the graph. On the next level for each metabolite all reactions are found. This continues while the level threshold is reached. As a result we receive a supergraph which represents all possible metabolites and reactions that are connected in one network.

Connectivity analysis of metabolites

To analyze the features of solution space of pathways constructed from the reactions available in the KEGG database and the necessary time for its construction practical experiments with *SAnalyzer* are performed. The performance of the presented algorithm is tested on the server computer with four Intel Xenon CPU and 32 GB RAM installed.

In this example the connectivity analysis options are shown for metabolites that can be used like a product or substrate. The time of graph construction algorithm, explored reactions and metabolites count for each construction level are shown. The data are shown for six metabolites – 2,3-Butanediol (KEGGID:C03046), *n*-Butanol(KEGGID:C06142), caprolactam(KEGGID:C06593), lactose (KEGGID:C00243), penicillin (KEGGID:C00395) and cellulose (KEGGID: C00395). 2,3-Butanediol is a chemical compound that is produced by a variety of microorganisms. It can be used in synthetic rubber producing and in gas chromatography. We can analyze producing pathways of this metabolite [27]. n-Butanol is alcohol, occurs naturally as a minor product of the fermentation of sugars and is present in many foods and beverages. The largest use of n-butanol is as an industrial intermediate, particularly for the manufacture of butyl acetate. We can analyze producing and consuming pathways of n-butanol[28]. Caprolactam is an organic compound and it is the precursor to Nylon 6, a widely used synthetic polymer[29]. We can analyze producing pathways of caprolactam. Lactose is sugar derived from galactose and glucose that is found most notably in milk. Lactose is not fermented by yeast during brewing. Another major use of lactose is in the pharmaceutical industry by adding it to pills as filler because of its physical properties. We can analyze consuming pathways of lactose to produce other substances [30]. Penicillin is a group of antibiotics, they are the first drugs that were effective against many previously serious diseases, such as syphilis, and infections caused by staphylococci and streptococci [31].

Cellulose is an organic compound, it is the structural component of the primary cell wall of green plants and many forms of algae. Converting cellulose from energy crops into biofuels such as cellulosic ethanol is under investigation as an alternative fuel source [32].

Using *SAnalyzer* tool it is possible to build a graph that represents metabolic pathways for the given substrate and get a list of nodes of this graph. Supergraph with 15 levels is constructed for each metabolite starting from the input substrate.

In this paper we present some data that can be obtained using *SAnalizer* tool. We show the count of connected metabolites on 15 levels. We can see similarities between all test cases. More additional data can be obtained using *SAnalyzer* scripts.

Results and discussion

For each test case the execution time, number of new added metabolites per level, number of added reactions per level and processing time per level have been measured. After 15 iterations more than 5000 metabolites are involved in the pathway tree for each test case (Fig. 1) from all 17000

metabolites kept in the KEGG database. In case of n-Butanol and caprolactam new metabolites are added slower than in other examples.

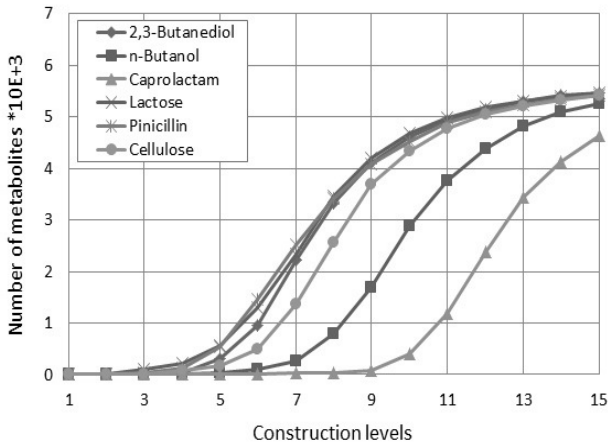

Fig. 1. **Number of metabolites in the constructed graph**

For caprolactam there are less than 5000 metabolites added in the constructed graph. The connectivity of caprolactam indicates that there are a small number of reactions to reach metabolites which are highly interconnected with other ones. Lactose in contrast demonstrates very fast growth of the metabolites involved in the pathway tree. All example cases show similar behavior with exponential growth on the number of new metabolites. For most metabolites it starts after 5 construction levels and after 10 iterations the number of new metabolites and reactions is increasing slowly. We can conclude that most of the reachable metabolites from KEGG are involved.

The execution time for each construction level (Fig. 2) demonstrates similar behavior for all 6 figures. Execution of 15 level graph construction takes about 10000 seconds that is about 3 hours for each metabolite. It correlates with new metabolite count graphs.

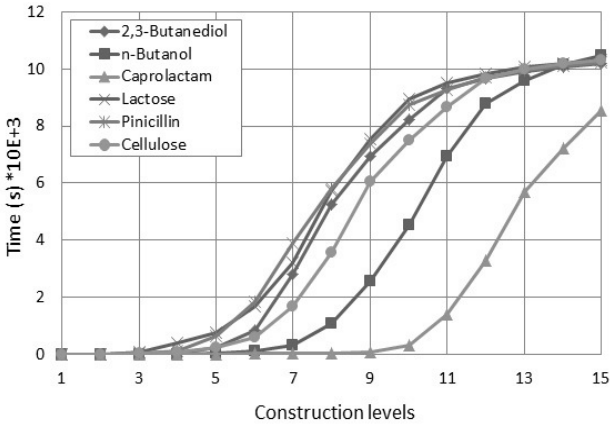

Fig. 2. **Runtimes of graph construction levels**

The number of newly included metabolites and reactions for each new level and time spent for each level demonstrate very clear peaks of the number of newly included reactions and metabolites in the area of 5-th till 11-th level (Fig. 3). Also it is possible to get the nodes and edges count of the constructed graph. After 15 iterations the nodes and edges count in all three cases is similar – about 14000 unique nodes (metabolites and reactions) and 31000 edges. A similar situation is for the explored reactions and metabolites count during supergraph construction.

This example does not provide complete information about the constructed graph metabolites and reactions. These data depend on the quality of KEGG data because for some reactions and metabolites the necessary data were not available. The achieved result shows the direction for future research in building new metabolic pathways.

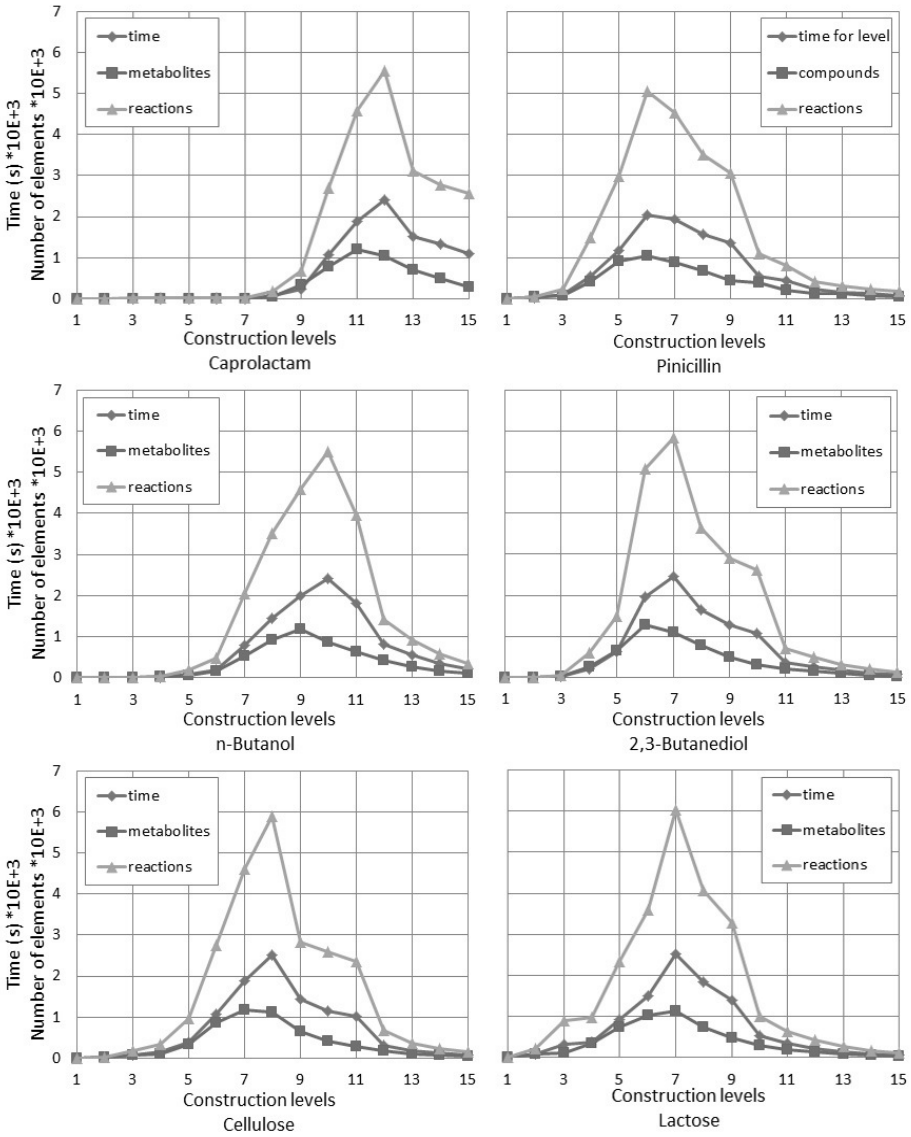

Fig. 3. **Runtime in seconds and the number of newly included elements (reactions and metabolites) - discovered in each construction level**

Conclusions

1. In this paper we present an application of *SAnalyzer* tool algorithms for possible metabolic pathways construction. Six structured supergraphs are built taking the initial substrate as the starting point and extending the graph by reactions from KEGG database.
2. This approach can be used to identify novel, non-native synthesis pathways for metabolite production.
3. It allows building a computer model of metabolic pathways from an initial compound to a product compound that potentially can be implemented into a living organism.
4. These data can be useful not only for analysis of biodegradation or producing new biological substances, but also for investigation of chemical transformations or chemical producing. One of the possible applications is recycling -a process that is using waste to produce potentially useful materials.
5. Experiments with six chemical substances are performed. The execution time, number of added metabolites per level, number of added reactions per level and processing time per level demonstrate similar behavior showing peaks in all the mentioned parameters around the level five till eleven. No significant further expansion of the supergraph after level thirteen is possible as most of the possible metabolites are already involved in the supergraph.

6. Additionally *SAnalyzer* functionality can be used to: construct possible pathways; examine how metabolites can be connected to each other, test different options of metabolic paths and show how biological data are connected related to KEGG database knowledge.

References

1. Stephanopoulos, G., Arisitidou, A., Nielsen, J. Metabolic engineering. Academic Press, San Diego, 1998. 707 p.
2. Rocha, I., Maia, P., Evangelista, P. etc. OptFlux: an open-source software platform for in silico metabolic engineering. BMC systems biology 4, 2010, p. 45.
3. Blazeck, J., Alper, H. S.. Systems metabolic engineering: Genome-scale models and beyond. Biotechnological Journal, vol. 5(7), 2010,pp. 647-659.
4. Trinh, C. T., Srienc, F. Metabolic engineering of Escherichia coli for efficient conversion of glycerol to ethanol. Applied and environmental microbiology, vol. 75(21),2009, pp. 6696-6705.
5. Trinh, C. T., Unrean, P., Srienc, F. Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. Applied and environmental microbiology, vol. 74(12), 2008, pp. 3634-3643.
6. Nielsen, J., Keasling, J.D. Synergies between synthetic biology and metabolic engineering. Nature biotechnology, vol. 29, 2011, pp. 693-695.
7. O'Malley, M. a, Powell, A., Davies, J.F. etc. Knowledge-making distinctions in synthetic biology. BioEssays, vol. 30, 2008, pp. 57-65.
8. Heinemann, M., Sauer, U. Systems biology of microbial metabolism. Current opinion in microbiology, vol. 13(3),2010, pp. 337-343.
9. Palsson, B. Ø. Systems Biology: Properties of Reconstructed Networks. New York: Cambridge University Press,2006, p. 322.
10. Thiele, I., Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols, vol.5(1), 2010, pp. 93-121.
11. Kostromins, A., Stalidzans, E. Paint4Net: COBRA Toolbox extension for visualization of stoichiometric models of metabolism. Biosystems, vol. 109(2),2012, pp. 233-239.
12. Rubina, T., Stalidzans, E.Topological features and parameters of Biochemical Network Structures. International Industrial Simulation Conference, 2010, pp. 228-236.
13. Schellenberger, J., Que, R., Fleming, R. M. T.etc. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols, vol. 6(9), 2011, pp.1290-1307.
14. Schellenberger, J., Park, J.O., Conrad, T.M. etc. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC bioinformatics, vol. 11, 2010, p. 213.
15. Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A. EcoCyc: a comprehensive database of Escherichia coli biology. Nucleic acids research, vol. 39, 2011, pp. D583-D590.
16. Karp, P.D., Ouzounis, C. a, Moore-Kochlacs, C. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic acids research, vol. 33, 2005, pp. 6083-6089.
17. Overbeek, R., Begley, T., Butler, R.M. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res., vol. 33(17), 2005, pp. 5691-5702.
18. Mednis, M., Aurich, M. K. Application of string similarity ratio and edit distance in automatic metabolite reconciliation comparing reconstructions and models. Biosystems and Information technology, vol. 1(1), 2012, pp. 14-18.
19. Mednis, M., Rove, Z., Galvanauskas, V. ModeRator - a software tool for comparison of stoichiometric models. 7th IEEE International Symposium on Applied Computational Intelligence and Informatics,2012, pp. 97-100.
20. Trinh, C.T., Wlaschin, A., Srienc, F. Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Applied microbiology and biotechnology, vol. 81, 2009, pp. 813-826.
21. Unrean, P., Trinh, C.T., Srienc, F. Rational design and construction of an efficient E. coli for production of diapolycopendioic acid. Metabolic engineering, vol. 12, 2010, pp.112-122.

22. Pentjuss, A., Odzina, I., Kostromins, A. etc. Biotechnological potential of respiring Zymomonasmobilis: a stoichiometric analysis of its central metabolism. Journal of biotechnology in press., 2013.
23. Yousofshahi, M., Lee, K., Hassoun,S. Probabilistic pathway construction, Metabolic Engineering, vol. 13, 2011, pp. 435-444.
24. Meitalovs, J. Software tool for probabilistic metabolic pathways construction. IEEE International Symposium on Computational Intelligence and Informatics,November 20-22, Budapest, Hungary, 2012, pp. 405-408.
25. Meitalovs. J. Analysis of synthetic metabolic pathways solution space. ICSSE 2013 IEEE International Conference on System Science and Engineering. July 4-6, Budapest, Hungary,2013. Unpublished.
26. Kanehisa, M., Goto, S., Sato, Y. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research, vol. 40, 2012, pp. D109-D114.
27. Voloch, M.,Jansen, N.B., Ladisch, M.R. etc. Fermentation Derived 2,3-Butanediol. Comprehensive Biotechnology, Pergamon Press Ltd., vol. 2(3), 1986, p. 933.
28. Hazelwood, L., Daran, J.M., Maris,A. etc. The Ehrlich pathway for fusel alcohol production: a century of research on Saccharomyces cerevisiae metabolism. Appl. Environ. Microbiol. vol. 74 (8), pp. 2259-2266.
29. Ritz, J., Fuchs, H., Kieczka, H., Moran, W. C. Caprolactam. Ullmann's Encyclopedia of Industrial Chemistry, 2011.
30. Linko, P., Lactose and Lactitol, Nutritive Sweeteners, London & New Jersey: Applied Science Publishers, 1982, pp. 109-132.
31. Garrod, L. P. Relative Antibacterial Activity of Three Penicillins. British Medical Journal, vol.1 (5172), 1960, pp. 527-529.
32. Holt-Gimenez, E. Biofuels: Myths of the AgrofuelsTransition.Backgrounder. Institute for Food and Development Policy, Oakland, CA., 2007, pp. 13:2.