Baltic J. Modern Computing, Vol. 3 (2015), No. 3, 164-178

Impact of Thermodynamic Constraint to the Solution Space of Metabolic Pathway Design Using *sAnalyzer* Tool

Jurijs MEITALOVS, Egils STALIDZANS

Latvia University of Agriculture, Faculty of Information Technologies, Liela iela 2, LV-3001, Jelgava, Latvia

jurijs.meitalovs@gmail.com, egils.stalidzans@gmail.com

Abstract: One of most important stages of metabolic engineering is organism design. One of methods that allows process design is modeling or *in silico* approach. Application of biological models instead of biological experiments is cheaper and does not require specific resources. *sAnalyzer* tool uses reactions data to construct new metabolic pathways and predict its feasibility to transform one metabolite to another one using particular chassis organism represented by a model. *sAnalyzer* searches possible reactions connections in full solution space of all accessible reactions. One of reaction parameters used to predict its flow direction is reaction thermodynamic data - Gibbs free energy (ΔrG) value, which shows how much energy is needed for reaction to flow in specific direction. *sAnalyzer* tool demonstrates how thermodynamic thresholds of ΔrG value impacts possible solution space of metabolites transformation pathways on the example of using glycerol as substrate for bacteria *Z.mobilis* model. The solutions are ranked by flux over the pathway of interest, biomass production ability and integrated parameters derived from involved metabolites and reactions.

Keywords: metabolic pathways analysis, thermodynamic data, *sAnalyzer* tool, solution space

1. Introduction

Metabolic engineering of microorganisms has increasing role in biotechnology (Patil, Akesson and Nielsen, 2004; Keasling, 2010). Metabolic engineering is closely related to other concepts in biotechnology as synthetic biology (Nielsen and Keasling, 2011) and cell factories. Modeling is important tool of metabolic engineering standing for mechanistic explanation and optimization of potential solutions (Blazeck and Alper, 2010) to produce the desired product from potential substrate introducing enzymes and regulatory elements from other species.

The aim is to determine how to produce the specified substance by introducing as small as possible number of new reactions into chassis microorganism. One way to do that is looking for metabolites transformation pathways in biological databases. Another way is to use computation tools that are able to find reactions connecting two

metabolites – reaction highway (Blum and Kohlbacher, 2008; Finley, Broadbelt and Hatzimanikatis, 2009; Ranganathan and Maranas, 2010; Rodrigo, Carrera, Prather and Jaramillo, 2008).

The same situation is when the substrate and the product are known but the most appropriate chassis organism in unknown (Fig. 1.). It is important to solve such input-output biotechnological task to find possible pathways of metabolite production.

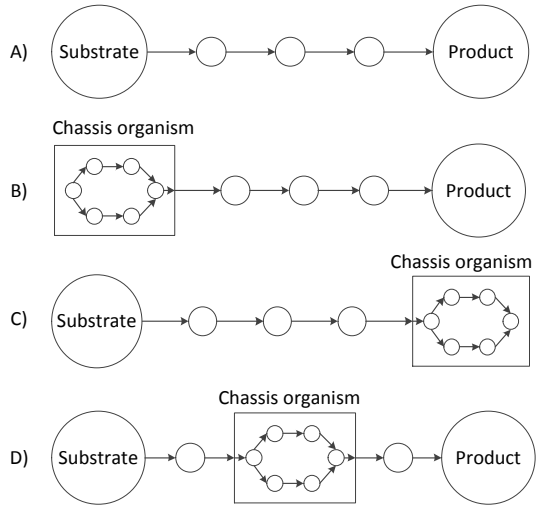

Fig. 1. The example of substrate – organism – product transformation pathways

Often the choice of organism and search of metabolic pathways happens in the experimental way or based on experimental data and organism-specific knowledge. Such approach does not guarantee the optimal result because just subsets of possible pathways are tested because of limited idea generation and evaluation resources.

There are also attempts to automatically analyze large number of possible combinations. Different are proposed by dedicated toolboxes (Hatzimanikatis et al., 2005; Blum and Kohlbacher, 2008; Meitalovs, 2012; Ranganathan and Maranas, 2010) without application of model of metabolic network.

Different metabolic pathways are analyzed in this study by the metabolic pathways construction and analysis computational tool *sAnalyzer* (Meitalovs and Stalidzans, 2013). The tool *sAnalyzer* can be used to find all possible specific metabolite related pathways of production or consumption for a specific chassis microorganism. Paper demonstrates application of *sAnalyzer* for search of metabolic engineering relevant pathways for glycerol conversion to ethanol by bacteria *Zymomonas mobilis*. Metabolites connection space analysis is performed for glycerol as a substrate using *Z.mobilis* metabolic model (Pentjuss et al., 2013) as a chassis representation.

2. Materials and methods

2.1. Algorithm

sAnalyzer is developed as a set of algorithms in Matlab environment. All algorithms are divided to three main functional modules executing separate steps described below. All experiments were performed using a computer with Intel i5 2.5GHz processor and Matlab 2013a installed.

At the **first** step *sAnalyzer* creates connection matrix of all possible transformation pathways of a specific metabolite and represents it in a way of graph. In such graph all reactions and metabolites are represent by nodes and metabolites participating in the particular reactions. Metabolites are connected as edges with the corresponding reaction. It is possible to configure parameters that determine connections matrix size of metabolic pathways: objective metabolite, maximum number of reactions in the transformation pathway, transformation flow direction – a substrate or a product, minimum biomass production (if included in the model), metabolites and reactions that should be skipped in rout construction and the thermodynamic value threshold of reactions.

Thermodynamic data of reactions is used to determine reactions that can happen only in one direction because of their thermodynamic properties (Fleming and Thiele, 2011) taking into account the stoichiometrically weighed sum of transformed Gibbs energy of products and substrates (Alberty, 2002; Alberty, 2003). To get the thermodynamic data of particular reactions the online calculator *eQuilibrator* is used.

sAnalyzer uses the change of standard Gibbs free energy of a given chemical reaction (Δ_rG) as threshold value to determine possible direction of the reaction flow. The change of standard Gibbs free energy shows how much energy is required to drive a particular biochemical reaction and in which direction the reaction will flow under particular cellular conditions. ΔG is a measure for the change of a system's free energy at constant pressure and temperature (Flamholz, Noor, Bar-Even and Milo, 2012).

If the module of thermodynamic value of reaction is below threshold value it is assumed that reaction is bidirectional (reversible). If thermodynamic value is above the threshold – the reaction is considered to have one direction.

For example, if threshold value is 30kJ/mol and reaction Δ_rG value is 24kJ/mol it is assumed that reaction can easily happen in both directions. In case when reaction Δ_rG value is 34kJ/mol or -34kJ/mol - reaction can flow only in one direction. The threshold value is adjustable.

Setting lower thermodynamic threshold value increases the number reactions to be evaluated on reaction direction. That reduces the total number of reactions in the possible solution space. Increase of the thermodynamic threshold increases the possible solution space but decreases the probability that the chosen reaction has expected direction.

At the **second** step *sAnalyzer* searches all possible pathways in the connection matrix that can connect the specific metabolite with a metabolite in the chassis microorganism model. It uses all paths *Depth-Firsts search* traversing algorithm to find all possible transformation pathways connected with the model in the connection matrix represented as a graph (Fig. 2).

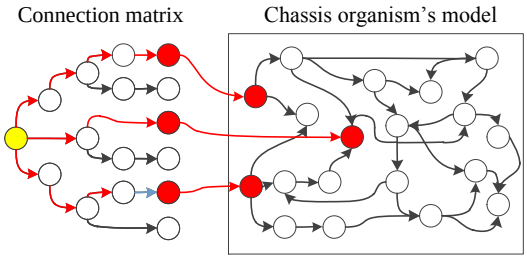

Fig. 2. Schematic representation of all possible pathways between a connection matrix and a chassis model

At the **third** step *sAnalyzer* analyzes earlier generated metabolic pathways to determine if the analyzed pathway is meaningful in biological context. Analysis algorithm uses the following parameters for ranking of solutions with weight coefficients:

1. Flux balance analysis data of analyzed pathway.
2. Flux balance analysis data of biomass reaction (or any another reaction marked as objective reaction).
3. Metabolites weight coefficient is calculated using linear interpolation method based on:
   - the total number of metabolites in the analyzed pathway;
   - the number of common metabolites between chassis model and the analyzed pathway;
   - the number of added metabolites to the model.
4. Reactions weight coefficient is calculated using linear interpolation method based on:
   - the total number of reactions in the analyzed pathway;
   - the number of common reactions between chassis model and the analyzed pathway;
   - the number of reactions with known thermodynamic data (if thermodynamic data is used);
   - the number of stoichiometrically unbalanced reactions in the pathway of interest.

It is possible to adjust weight coefficients of each parameter to make it more or less significant or assign no impact of the parameter. *sAnalyzer* allows to use two analysis methods: **separate** - where results can be sorted based on each separate parameter, and **combined** – where al parameters are combined into one weight using complex parameters weighting method and results are sorted by this value.

In the result *sAnalyzer* returns the list of metabolites and corresponding reactions which allows transforming the objective metabolite into another one, which is presented in the chassis metabolic model and the organism is able to growth.

2.2. Data exchange

sAnalyzer works with *COBRA* SBML metabolic model format. *COBRA* allows adding additional fields into model to complete model information, e.g. reactions or metabolites common used identifiers that may be missing in the original model. *sAnalyzer* uses identifiers from "Kyoto Encyclopedia of Genes and Genomes" online database's (KEGGID) (Ogata et al., 1999). *sAnalyzer* uses this field to identify and connect metabolites in transformation pathways.

Usually, in SBML standard (Hucka et al., 2003) models KEGGID information is not available and instead of manual work *sAnalyzer* allows filling KEGGID data automatically using KEGG database's search REST API. If it is not possible to find this information *sAnalyzer* leaves this field blank. Sometimes it is impossible to find necessary information using KEGG API and manual review and KEGGIDs search is needed. That is relatively simple to do for a small models, e.g. core metabolism models like *Z. mobilis* core model (Pentjuss et al., 2013), which contains 80 reactions while it is time consuming for models with thousands of metabolites.

2.3. Task setting

Recent publication (Meitalovs and Stalidzans, 2013) presents the number of discovered metabolites in the connection matrix dependent on the maximum number of reactions in the transformation pathways for different metabolites. It is done using *sAnalyzer* connection matrix building module.

In this work we present possible solution space for glycerol transformation pathways for *Z. mobilis* core model (Pentjuss et al., 2013). Connection matrix for glycerol with maximum 10 reactions in the pathway was built. The list of metabolites that cannot be used in the main atom transfer was created (ATP, ADP, NADH, NAD, CoA, H, NADPH, H20, NADP, O2, Phosphate, CO2, PPi, NH3, UDP, H2O2, SAM). Experiments were done using thermodynamic threshold (ThD) ΔrG' 5 (kJ/mol) and without thermodynamic threshold. *sAnalyzer* uses thermodynamic data from online service *eQuilibrator*. This service allows configuring environment variables and getting thermodynamic data for the reaction in this given environment (Flamholz et al., 2012). Thermodynamic threshold is used to filter reactions with opposite flux direction using certain cellular conditions. Without using thermodynamic threshold all reactions are used in construction metabolites connection matrix – it is assumed that all reactions are reversible and can flow in both directions. The same situation is when *eQuilibrator* does not contains thermodynamic data for reaction – reaction is assumed to be reversible.

The constructed connection matrix can be useful to guess the scale of potential connection space for specific metabolite – how many biological metabolic pathways can transform it to other substances. It shows as well how many metabolites and reactions are connected in one network.

3. Results and discussion

3.1. Impact of maximum number of reactions to metabolite connections space

sAnalyzer can build a connection matrix of a specific metabolite transformation pathways and represents them as a connection graph. Example graph for glycerol transformation, limited by maximum one reaction in transformation pathway and thermodynamic threshold 30 kJ/mol is demonstrated in Figure 3 using KEGG identifiers of metabolites and reactions. Initial metabolite is glycerol (C00116).

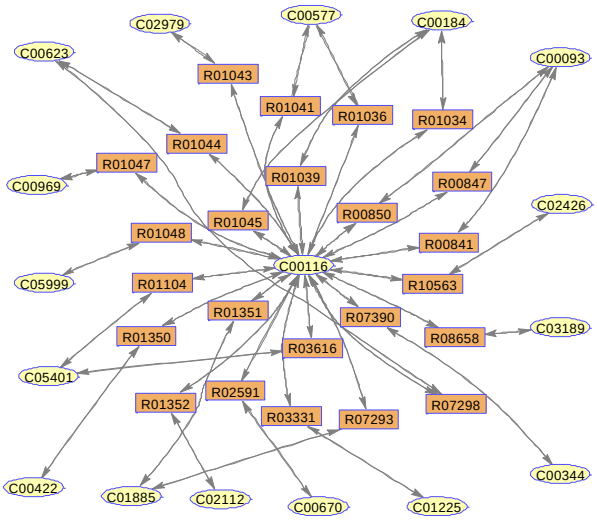

Fig 3. Visualized stoichiometric matrix constructed from 41 nodes - 24 reactions (first letter is "R") and 17 metabolites (first letter is "C") and 48 edges – connections between nodes

The graph (Fig. 3) includes small number of metabolites and reactions. To solve biotechnological problem usually it is necessary to analyze larger amount of possible reactions. Figure 4 shows number of explored metabolites and reactions for construction of connection matrix, limited by maximum number of reactions in the pathway from 1 to 10 using thermodynamic threshold 5kJ/mol and without it. The biggest number of new metabolites and reactions is explored between connection levels 4 and 7. After level 7 the increase of newly explored metabolites and reactions slows down.

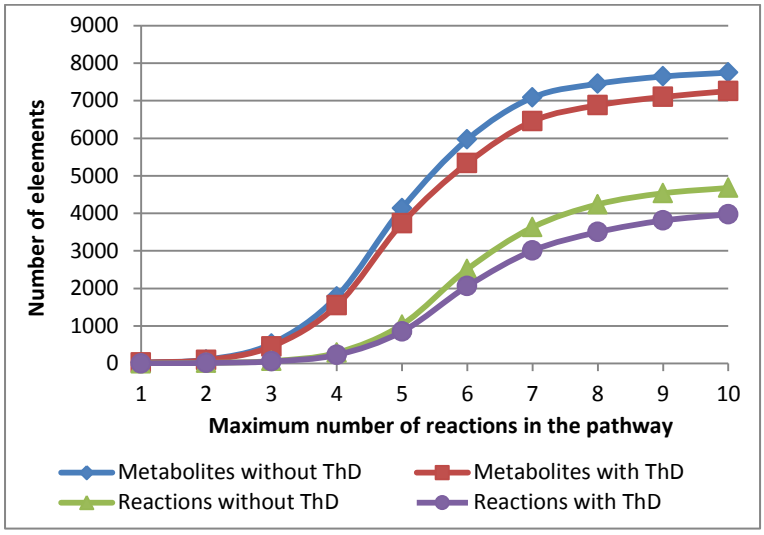

Fig. 4. The number of connected metabolites and reactions count depending on maximum number of reactions in pathway from 1 to 10

Implementation of thermodynamic data (ThD) reduces the number of feasible reactions because of constraining of many bidirectional reactions to a single direction reaction in the given environment. The Richard's type curve shows limitation of reactions and metabolites as most of them have been already involved in the connection matrix. It shows that reasonable maximal number of reactions in pathway used in the future analysis is 7, as further increase of involved metabolites and reactions is small while computational load grows very fast.

3.2. Building a connection matrix of glycerol transformation pathways

To show the impact of thermodynamic data to connection matrix size four experiments were done with limiting maximum reactions count in pathways by 1, 3, 5 and 7 reactions. Each experiment was done with different thermodynamic threshold limit – without thermodynamic threshold (No Thd) - all reactions where used to build metabolic pathway, thermodynamic thresholds 30 kJ/mol (30 ThD) and 5 kJ/mol (5 Thd). If no thermodynamic data is available for reaction – it is assumed that it can happen in both directions.

Thermodynamic (ΔrG) data source is downloaded as CSV file from *eQuilibrator* for cellular parameters: pH level 7.0, ionic strength 0.1mM and temperature 298.15K. The file contains data about 8336 reactions and 3889 of them do not have ΔrG value. As it was seen on the Figure 4 threshold 5 kJ/mol filtered many reactions that happen in opposite direction than metabolic transformation of glycerol should take place.

12 connection matrixes were built (Fig. 5) to analyze connection matrix size of each experiment and configuration. Different thermodynamic threshold (ThD) is used to limit

reactions direction. Y-axis is in a logarithmical scale. The largest connection matrix is built with maximum limitation of 7 reactions and without thermodynamic threshold.

Introduction of thermodynamic threshold 5 kJ/mol the number of connection nodes decrease by 2800 reactions and metabolites and by more than 3000 edges connecting reactions and metabolites.

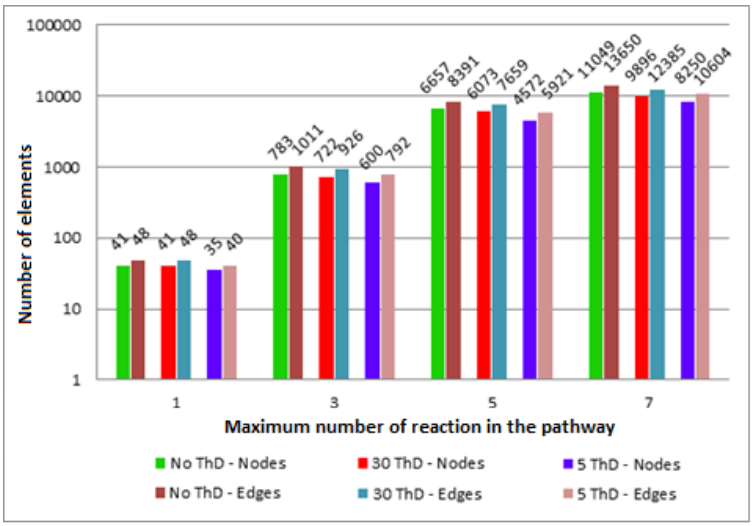

Fig. 5. Number of nodes and edges in graph of constructed connection matrix for maximum 1, 3, 5 and 7 reactions pathways

Building of such connection matrix takes comparatively small amount of time – the largest matrix can be built in one hour. Figure 6 shows time consumption for each experiment.

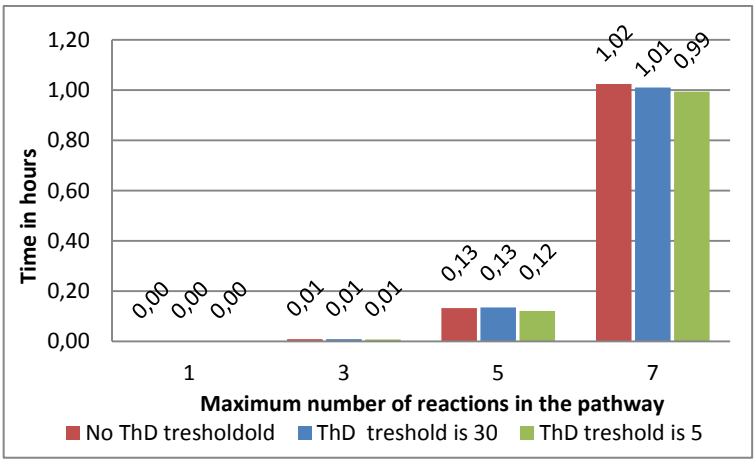

Fig. 6. Time spent on connection matrix construction for maximum 1, 3, 5 and 7 reactions pathways depending on reaction directionality limiting thermodynamic threshold (ThD)

3.3. Search of transformation pathways in connection matrix

In this step *sAnalyzer* generates all possible glycerol transformation pathways in the connection matrix built in previous step to any enabled metabolite in given *Z.mobilis* metabolic model. Figure 7 shows the number of found metabolic transformation pathways of glycerol and *Z.mobilis* metabolic model (Pentjuss et al., 2013). Y-axis is given in a logarithmical scale.

Application of small thermodynamic threshold (5 kJ/mol) reduces the number of found pathways for about 14 times compared to 30 kJ/mol. That shows that relatively small increase of number of nodes and edges in the connection matrix causes a combinatorial explosion of the number of possible connection pathways.

Therefore, many pathways with irreversible reactions are rejected automatically thus shrinking the solution space.

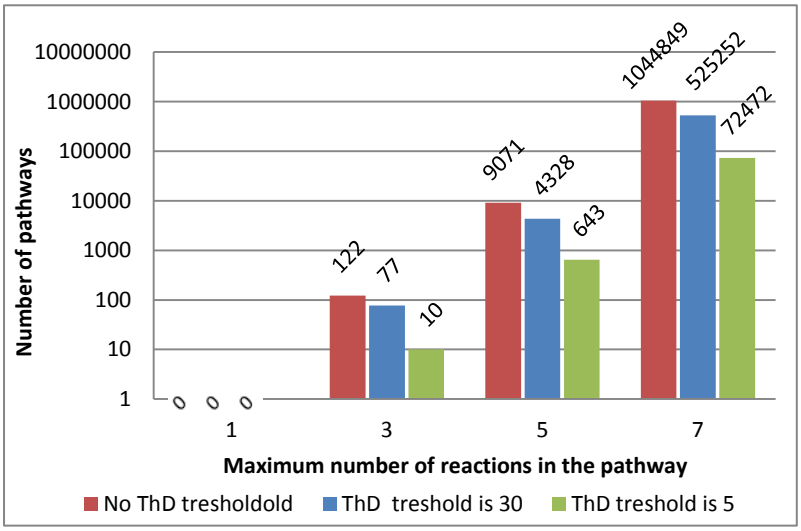

Fig. 7. Number of found metabolic pathways in constructed connection matrix between glycerol and *Z. mobilis* metabolic pathway

This process is time consuming for big connection matrixes. Application of thermodynamic data constraint causes rejection of 2800 connection nodes and more than 3000 edges decreasing the processing time of 7 reactions graph more than 5-fold (Fig. 8). At the same time the calculation time for up to 5 reactions is reduced about 2-fold.

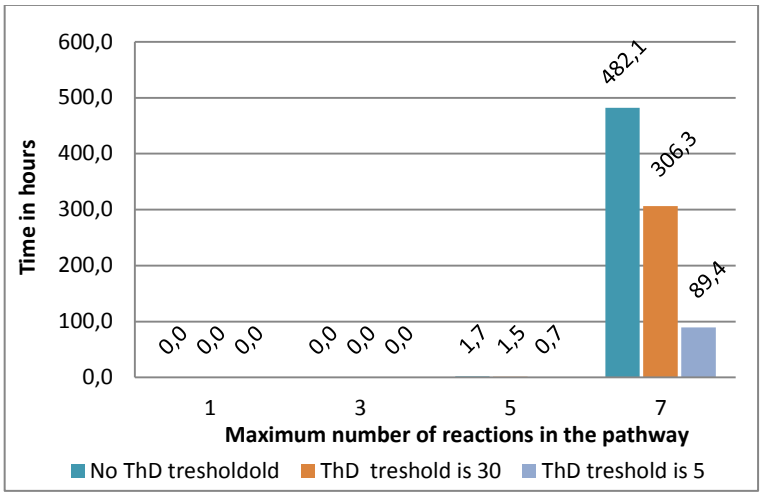

Fig. 8. Time spent on transformation pathways search for maximum 1, 3, 5 and 7 reactions pathways depending on the applied thermodynamic threshold (ThD)

3.4. Ranking of proposed pathways

During this step previously constructed pathways are ranked by their expected biotechnological applicability. Constructed pathways one by one are implemented into the model using COBRA toolbox functionality and FBA analysis is performed to determine the flow of metabolites via added pathway and via biomass reaction of the model. In that way the integration of newly added path is tested on its ability to function in a model of chassis organism. FBA results, reactions and metabolites parameters multiplied by weight coefficients are used to sort and rank the list of analyzed pathways by their potential of successful applicability. It is possible to change context information by weight coefficients to stress on particular features of the solution making some pathway parameters more important. It is possible to limit maximum number of top-ranked pathways returned to user for further analysis.

In the performed ranking the minimal required production of biomass and nonzero flux over the implemented pathway were requested as minimal condition for a pathway to stay in the implementation candidates list. Figure 9 shows analyzed and sorted number of transformation pathways that satisfy minimal preconditions of biomass production and nonzero flux of interest and could undergo manual analysis. The maximum number of top-ranked pathways is limited by 5000. There is a heavy decrease of number of acceptable pathways when they are implemented in the model compared to the corresponding number of pathways in Fig. 7.

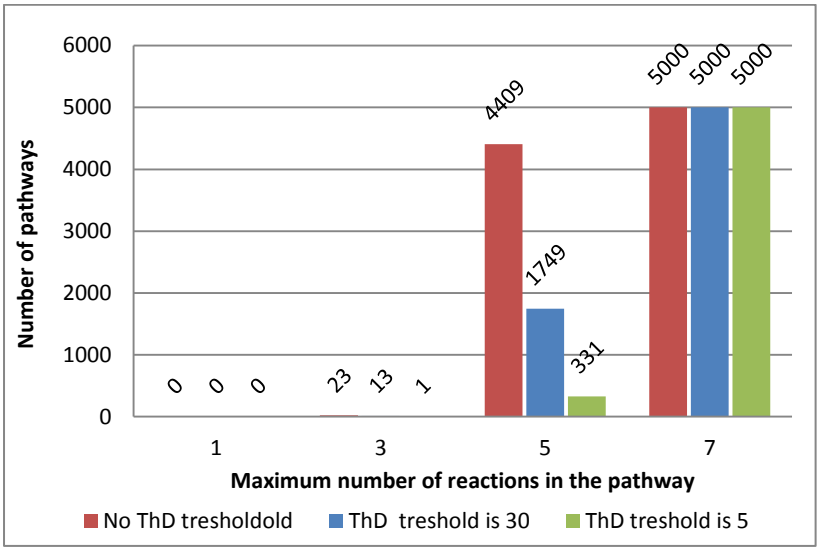

Fig. 9. Number of ranked metabolic pathways in constructed connection matrix that enable biomass production and nonzero flux through the introduced pathway

This process also is time consuming for big connection matrixes (Fig. 10). Different thermodynamic threshold (ThD) is used to limit reactions direction. The ranking for pathways with maximum 7 reactions without thermodynamic threshold compared to the same task with thermodynamic threshold 5 kJ/mol takes 12-fold more time. Comparing overall spent time to generate and to rank all possible pathways with and without thermodynamic data the time reduction is 7-fold.

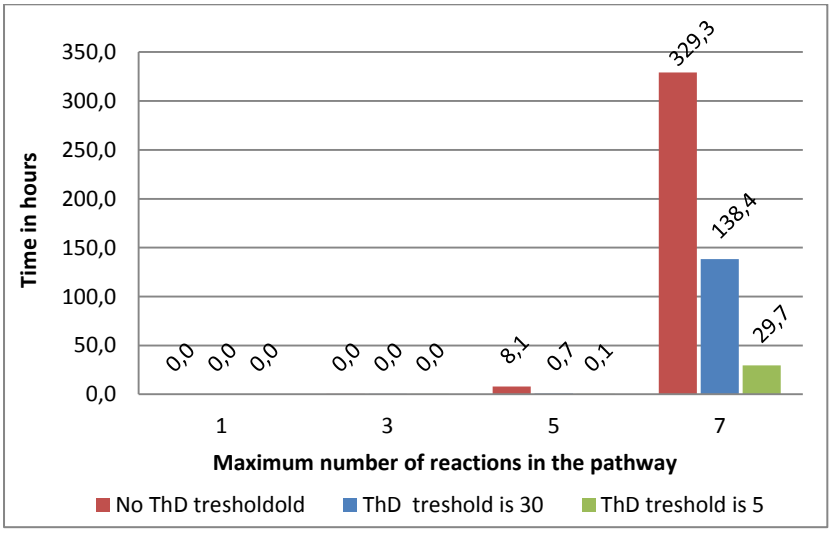

Fig. 10. Time spent on transformation pathways analysis for maximum 1, 3, 5 and 7 reactions pathways

For cases with relatively small maximum number of reactions in the pathway, e.g. 1 to 5, application of thermodynamic data do not deliver great savings of computational time as connection matrix is relatively small. On average ranking of 3000 pathways takes one hour.

3.5. Visualisation of pathways

After analysis it is possible to visualize chosen pathways. On the Figure 11-A 10 top-ranked pathways for experiment without thermodynamic data and 7 reactions in the pathway can be connected only to one metabolite in the model. Figure 11-B shows top 10 transformation pathways with reactions thermodynamic threshold 30 kJ/mol and 7 reactions in the pathway.

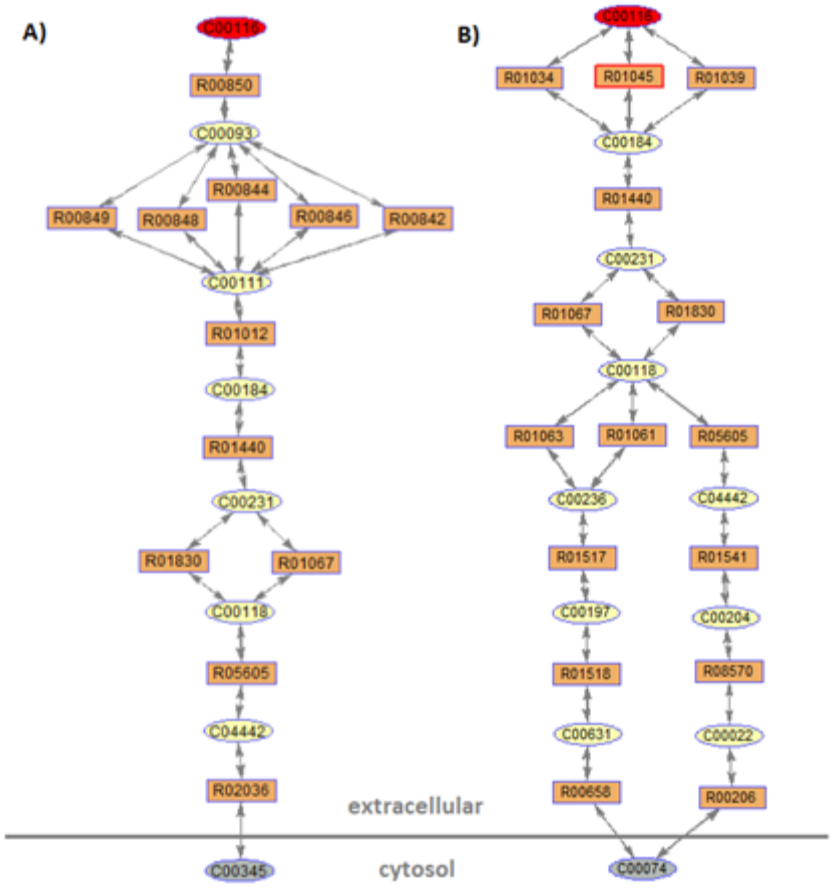

Fig. 11. Automatic generated representation of Top 10 analyzed pathways with maximum number of reactions 7. A) - pathways without thermodynamic threshold, B) - with thermodynamic threshold 30 kJ/mol (red – initial metabolite, grey – metabolite that is connected to the model)

Metabolite which connects analyzed pathway with the chassis model (*cytosol*) is different in both cases because one or several reactions in Figure 11-A cannot transform metabolites in particular direction if thermodynamic data are applied.

Figure 12 shows top 50 pathways for experiment with thermodynamic threshold 5 kJ/mol and maximum 5 reactions in the pathway. Many irreversible reactions with wrong reaction flow direction are filtered out and in the result pathway have only five connection metabolites with the chassis organism model in cytosol.

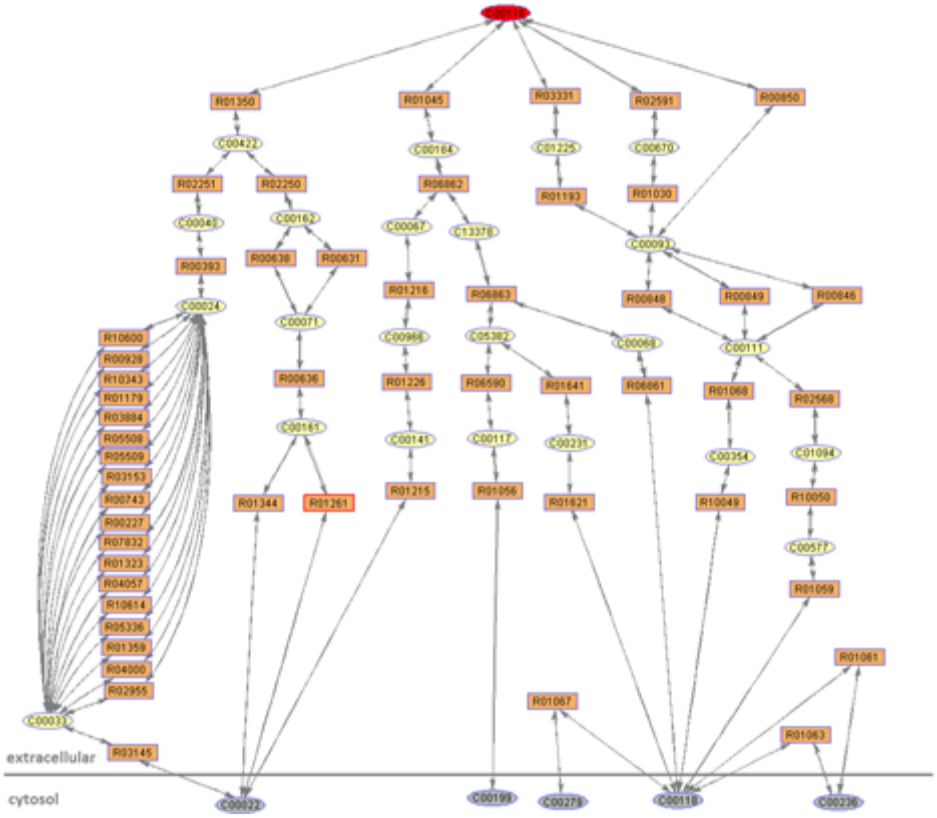

Fig. 12. Automatic generated representation of Top 50 analyzed pathways using thermodynamic threshold 5 kJ/mol and maximum number of reactions 5

The top-ranked reactions have bigger probability to be successfully implemented into real organism. Still important biological information is not taken into account and the most promising pathways should be manually examined by biologists before they are accepted for biological experiments. The proposed automation approach reduces manual work at the same time enabling scan of complete solution space automatically by a computer.

4. Conclusions

sAnalyzer software tool is a powerful instrument for automatic generation and ranking of metabolic pathways from a particular substrate to any metabolite in the stoichiometric model of chassis organism taking into account number of 1) metabolites in the pathway; 2) common metabolites in the pathway and model; 3) newly added metabolites; 4) reactions in the pathway; 5) reactions in the model; 6) reactions with known thermodynamic and 7) number of unbalanced reactions with different weight factors.

The number of possible metabolic pathways for utilization of a particular substrate rapidly increases proportionally to the maximal number reaction introduced.

The implementation of thermodynamic constraints heavily reduces both the solution space and computational time rejecting solutions that are not feasible because of thermodynamic limitations. All analysis steps can take long time for big size connection matrix, e.g. for presented glycerol transformation analysis without thermodynamic limits takes 812 hours, that is almost 34 days. Using thermodynamic threshold 5 kJ/mol all analysis steps take 120 hours or 5 days. Implementation of thermodynamic data reduces computation time about 5-fold.

Parallel computing can be implemented to reduce computational time at costs of increasing computational power – RAM and CPU speed. Matlab offers parallel computing (Parallel computing toolbox) options but behaviour of its implementation differs on different OS and Matlab versions.

Additionally to KEGG database, another data sources could be coupled with *sAnalyzer*.

References

Alberty, R. A. (2002). Thermodynamics of systems of biochemical reactions. *Journal of Theoretical Biology*, *215*, 491–501. doi:10.1006/jtbi.2001.2516

Alberty, R. A. (2003). *Thermodynamics of Biochemical Reactions* (p. 397). Hoboken, New Jersey: Wiley.

Blazeck, J. and Alper, H. S. (2010). Systems metabolic engineering: Genome-scale models and beyond. *Biotechnological Journal*, *5*(7), 647–659. doi:10.1002/biot.200900247.

Blum, T. and Kohlbacher, O. (2008). MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics (Oxford, England)*, *24*(18), 2108–9. doi:10.1093/bioinformatics/btn360

Finley, S. D., Broadbelt, L. J. and Hatzimanikatis, V. (2009). Computational framework for predictive biodegradation. *Biotechnology and Bioengineering*, *104*(6), 1086–97. doi:10.1002/bit.22489

Flamholz, A., Noor, E., Bar-Even, A. and Milo, R. (2012). eQuilibrator - The biochemical thermodynamics calculator. *Nucleic Acids Research*, *40*. doi:10.1093/nar/gkr874

Fleming, R. M. T. and Thiele, I. (2011). von Bertalanffy 1.0: a COBRA toolbox extension to thermodynamically constrain metabolic models. *Bioinformatics (Oxford, England)*, *27*(1), 142–3. doi:10.1093/bioinformatics/btq607

Hatzimanikatis, V., Li, C., Ionita, J. a, Henry, C. S., Jankowski, M. D. and Broadbelt, L. J. (2005). Exploring the diversity of complex metabolic networks. *Bioinformatics (Oxford, England)*, *21*(8), 1603–9. doi:10.1093/bioinformatics/bti213

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524–531. doi:10.1093/bioinformatics/btg015

Keasling, J. D. (2010). Manufacturing molecules through metabolic engineering. *Science (New York, N.Y.)*, *330*(6009), 1355–8. doi:10.1126/science.1193990

Meitalovs, J. (2012). Software tool for probabilistic metabolic pathways construction. *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)*, 405–408. doi:10.1109/CINTI.2012.6496800

Meitalovs, J. and Stalidzans, E. (2013). Analysis of synthetic metabolic pathways solution space. *2013 International Conference on System Science and Engineering (ICSSE)*, 183–187. doi:10.1109/ICSSE.2013.6614656

Nielsen, J. and Keasling, J. D. (2011). Synergies between synthetic biology and metabolic engineering. *Nature Biotechnology*, *29*(8), 693–5. doi:10.1038/nbt.1937

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *27*(1), 29–34. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148090&tool=pmcentrez&rendertype=abstract

Patil, K. R., Akesson, M. and Nielsen, J. (2004). Use of genome-scale microbial models for metabolic engineering. *Current Opinion in Biotechnology*, *15*(1), 64–9. doi:10.1016/j.copbio.2003.11.003

Pentjuss, A., Odzina, I., Kostromins, A., Fell, D. A., Stalidzans, E. and Kalnenieks, U. (2013). Biotechnological potential of respiring Zymomonas mobilis: A stoichiometric analysis of its central metabolism. *Journal of Biotechnology*, *165*, 1–10. doi:10.1016/j.jbiotec.2013.02.014

Ranganathan, S. and Maranas, C. D. (2010). Microbial 1-butanol production: Identification of non-native production routes and in silico engineering interventions. *Biotechnology Journal*, *5*(7), 716–25. doi:10.1002/biot.201000171

Rodrigo, G., Carrera, J., Prather, K. J. and Jaramillo, A. (2008). DESHARKY: Automatic design of metabolic pathways for optimal cell growth. *Bioinformatics*, *24*(21), 2554–2556. doi:10.1093/bioinformatics/btn471

Received April 16, 2015, accepted July 24, 2015