# Evaluating Pathway Enumeration Algorithms in Metabolic Engineering Case Studies

F. Liu, P. Vilaça, I. Rocha, and Migael Rocha

CEB/IBB, Universidade do Minho, Portugal

**Abstract.** The design of cell factories for the production of compounds involves the search for suitable heterologous pathways. Different strategies have been proposed to infer such pathways, but most are optimization approaches with specific objective functions, not suited to enumerate multiple pathways. In this work, we analyze two pathway enumeration algorithms based on graph representations: the Solution Structure Generation and the Find Path algorithms. Both are capable of enumerating exhaustively multiple pathways using network topology. We study their capabilities and limitations when designing novel heterologous pathways, by applying these methods on two case studies of synthetic metabolic engineering related to the production of butanol and vanillin.

## 1 Introduction

The quest for sustainable industries lead to an increased interest in Biotechnology. One of its key features is to re-engineer microbes to produce valuable compounds [5]. The development of cell factories is an iterative process involving steps as the search for suitable hosts and viable synthetic pathways. Heterologous pathways augment their capabilities to produce non native compounds. The definition of pathways allows to organize chemical reactions into set providing a coherent function, such as transforming a substrate to a target compound.

The constraint based modeling (CBM) approach is often adopted for *in silico* analysis of genome scale metabolic models (GSMM) not requiring kinetic information. The system is subjected to constraints such as reaction stoichiometry, reversibility and assumption of a pseudo-steady state, allowing the computation of a feasible flux space that characterizes the system. Flux Balance Analysis (FBA) is a popular method to determine the flux distribution that maximizes an objective (e.g. related to cellular growth) using linear programming [16].

Pathway optimization has been approached using different strategies. Regarding CBM, FBA was used to determine producible non native compounds [3] by merging GSMMs with large databases as KEGG, allowing to infer heterologous reactions. A limitation of FBA is the fact that it determines a single solution, while multiple optimal solutions exist. Furthermore, sub-optimal solutions may offer valuable information on alternative routes. On the other hand, Elementary Flux Modes (EFM) are defined as the minimal subsets of reactions to maintain steady state. However, their computation is restricted to small networks [14].

J. Sáez-Rodríguez et al. (eds.), *8th International Conference on Practical Appl. of Comput. Biol. & Bioinform. (PACBB 2014)*, Advances in Intelligent Systems and Computing 294,
DOI: 10.1007/978-3-319-07581-5_26, © Springer International Publishing Switzerland 2014

Figueiredo *et al* [6] propose an enumeration strategy to compute the $k$ shortest EFMs expanding the size of partially computable problems. Nonetheless, database size networks (e.g. KEGG or MetaCyc) still offer a great challenge for full EFM computation. The OptStrain algorithm [17] uses mixed integer linear optimization to obtain the pathway with the smallest number of heterologous reactions, but does not enumerate alternatives.

Other methods have applied standard graph methods, taking advantage of shortest path algorithms to infer the shortest pathway between two compounds [8,18]. This strategy can also be augmented by using shortest path enumerating methods, such as the $k$-shortest path algorithm [4]. A major problem with this strategy is that graph paths return linear routes between compounds, while in reality these may involve more compounds. Additionally, compounds represented as hubs in the network mislead the algorithms by shortening the paths since they connect many reactions. To circumvent this problem, weighting [8] or filtering methods [7] have been proposed to reroute the solutions.

An alternative is to use more complex representations. Hypergraphs or process graphs (which are directed bipartite graphs) are capable to model chemical reactions with higher detail. This allows to address the problem of multiple products and reactants, since edges connect to vertex sets instead of a single vertex. Process graphs were used by Friedler *et al* [9–11] in an exhaustive approach for decision mapping in synthesis processes, being later adapted for pathway identification [13]. The work of Carbonell *et al* [2] introduced an enumeration strategy to extract pathways using hypergraphs.

In this work, we analyze two existing algorithms for multiple pathway enumeration, the Solution Structure Generation (SSG) and the Find Path (FP), both based on set systems representations. These algorithms are implemented and tested with two case studies, regarding the production of butanol and vanillin, using the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae*, two model organisms for which there are available GSMMs. The results obtained by both are provided and discussed, being clear the need to introduce some improvements to allow the scalability of the methods.

## 2 Problem Definition

In a topological approach, a pathway extraction problem can be defined as a dependency problem. Thus, a reaction needs to be satisfied and satisfies metabolites (that are dependencies of other reactions), that correspond to reactants and products, respectively. Here, the notation used in the following is defined. Mostly, it is based on the axioms and algorithms presented in [9–11].

Networks will be composed only by metabolites and reactions. In this system, metabolites are the vertex entities, while reactions are represented by an ordered pair $\langle M_1, M_2\rangle$, that connects two disjoint sets of metabolites.

**Definition 1.** *(Reaction) A reaction is an ordered pair $\langle M_1, M_2\rangle$ of two disjoint sets of metabolites (i.e., $M_1 \cap M_2 = \emptyset$). The first set represents the reactants, while the second represents the products.*

**Definition 2.** *(Metabolic Network) A metabolic network $\Sigma$ is a pair composed by a set of metabolites $\Pi$ and a set of reactions $\Upsilon$.*

A reversible reaction $r$ is represented by including another entity $r'$, such that the metabolite sets are swapped. Additionally, a network $\Sigma' = \langle \Pi', \Upsilon' \rangle$ is defined as a subnetwork of $\Sigma\langle \Pi, \Upsilon \rangle$ if every element of $\Sigma'$ is contained in $\Sigma$ (i.e., $\Pi' \subseteq \Pi$ and $\Upsilon' \subseteq \Upsilon$), then $\Sigma' \subseteq \Sigma$.

A retrosynthetic metabolic problem can be defined as follows:

**Definition 3.** *(Retrosyntehtic Metabolic Problem) A retrosynthetic metabolic problem $\Gamma$ is defined by a triplet $\langle \Sigma, S, T \rangle$, where $\Sigma$ is a metabolic network that represents the search space, while $S$ and $T$ are two disjoint sets of metabolites (i.e, $S \cap T = \emptyset$) which are the constraints of the heterologous pathways. The set $S$ keeps the initial substrates (e.g., supplies or raw materials), while the set $T$ defines the target compounds of interest.*

An heterologous pathway is a set of reactions, in most cases a subnetwork of a larger network (defined as the search space), if it satisfies the following:

**Definition 4.** *(Heterologous Pathway) An heterologous pathway $\sigma$ of a synthetic problem $\Gamma$ is any network (or subnetwork), such that: a) the product set $T$ is included in $\langle M, R \rangle$, i.e., $T \subset M$ and b) for every metabolite $m$ in the subnetwork that is not included in the substrate sets of $\Gamma$ (i.e., $M - S$) there is a reaction $r$ in $R$ such that $m$ is a product of $R$.*

The heterologous pathway definition is not sufficient to guarantee that the solution is feasible, because it omits the stoichiometry of the reactions. Both algorithms addressed in this work do not take account this property for the computation of heterologous solutions. This eventually will lead to the computation of infeasible solutions that later can be verified by applying FBA.

# 3 Algorithms

## 3.1 Solution Structure Generation

The Solution Structure Generation (SSG) algorithm enumerates solutions of $\Gamma$ by recursively branching all possible combinations. This technique, denoted as decision mapping, can be described as follows: let $\Sigma'$ be a subnetwork such that condition a) verifies. Then, in order to fulfill condition b), the sub-problem $\Gamma'$ is solved producing the unsatisfied metabolites in $\Sigma'$. Let $\Sigma = \langle T, \emptyset \rangle$ be a network containing $T$ and no reactions, then a) trivially verifies. Then, $\wp$(producers of $t$), $t \in T$ (where $\wp(X)$ denotes the power set of $X$) are candidates for partial solutions of $\Gamma$, since if solutions of $\Gamma$ exists, then at least one element of $\wp$ eventually must be present in one or more solutions of $\Gamma$. Recursively, we solve the sub-problem $\Gamma'$, with the new target set $T' = R - S - M$, where $R$ is the set of reactants of the newly introduced reactions (minus the initial set $S$ and producible metabolites in the partial solution), until eventually either there are no possible reactions to add, and this implies that we have reached a dead end that happens when we pick a producer of $T$ that does not belong to any solution, or $T = \emptyset$ which implies that we achieved a solution.

There are several limitations of the SSG method. The first is the high amount of memory that is required to compute power sets which grows exponentially with the number of elements ($2^n$). Additionally, this generates an extensive amount of possible combinations. If the network is not pruned, meaning that the network contains reactions that do not belong to any solution, then the algorithm may contain branches that return no solutions and, depending the depth of these branches, this increases severely the computation time to obtain solutions. Friedler *et al* [10] proposed a polynomial algorithm to prune process graphs to remove all reactions that might exhibit this behavior. Because of these limitations, in this work, some modifications were implemented to the original algorithm. Given space constraints, the full algorithms including these changes are fully given and explained in supplementary material that is available in `http://darwin.di.uminho.pt/pacbb14-liu`.

### 3.2 Find Path

The Find Path (FP) algorithm proposed by Carbonell *et al* [2] enumerates pathways by using hypergraphs. In a metabolic context, both hypergraphs and process graphs are much similar. A solution of the FP algorithm is defined as a *hyperpath* $P$, which is an hypergraph (usually a subgraph) where the hyperarcs (reactions) can be ordered as $r_1, r_2, \ldots, r_m$ such that $r_i$ is dependent only on the substrates in $S$ and the products of the previous reactions. This is computed with a subroutine, Find All [2], that sorts the entire network satisfying this condition. Additionally, reactions that cannot be satisfied are removed.

Not all pathways can be expressed by the definition of an hyperpath [2]. Lets consider for instance co-factor metabolites $m_a$ and $m_b$. Usually, these metabolites are both present in a single reaction $r = \langle M_1, M_2 \rangle$ where $m_a \in M_1$ and $m_b \in M_2$ or vice versa. These reactions can be satisfied by each other in a way where there is an $r' = \langle M'_1, M'_2 \rangle$ where $m_b \in M'_1$ and $m_a \in M'_2$. Therefore, it is impossible to sort an hyperpath if neither $m_a$ or $m_b$ are included in $S$. Examples of these metabolites are ATP/ADP, NADH/NAD, etc. Fortunately, if assuming $S$ to be an organism chassis, these metabolites are usually include in $S$. However, this does not guarantee that other more complex cycles do not exist.

This issue enables the generation of redundant solutions. Let $\Gamma = \langle \Sigma, \{s_0\}, \{t_0\} \rangle$ be a retrosynthetic problem, assuming that a) an heterologous pathway $\Sigma' \subset \Sigma$ exists from $s_0$ to $t_0$, such that b) $r, r' \in \Sigma'$ where $r = \langle \{m_0, p_0\}, \{m_1, p_1\} \rangle$ and $r' = \langle \{m_1, p_1\}, \{m_2, p_0\} \rangle$. The FP algorithm can only identify such pathway if $\Gamma' = \langle \Sigma, \{s_0\}, \{p_0, m_0\} \rangle$ is feasible. Since $r, r'$ satisfy the metabolites $p_0, p_1$ of each other (i.e., $r + r' = \langle \{m_0\}, \{m_2\} \rangle$) this implies that any effort to produce $p_0$ in $\Gamma'$ is unnecessary and every solution that b) verifies may contain multiple redundant solutions (the reactions included in the solutions are unique but in steady state they are redundant).

In this work, to extend the capabilities of the FP algorithm a modification was implemented in the Minimize subroutine (see supplementary material in `http://darwin.di.uminho.pt/pacbb14-liu`. The redundancy problem still remains an open topic for further improvement.

## 4 Experiments and Results

### 4.1 Case Studies

The algorithms were tested by applying two case studies of synthetic metabolic engineering. The first example is the production of 1-butanol using *E. coli* [1], while the second concerns vanillin synthesis using *S. cerevisiae* [12]. Both algorithms (i.e., SSG and FP) are applied using the set of compounds in the KEGG Ligand and MetaCyc databases as the chemical search space. Additionally, to integrate and test the obtained solutions *in silico*, a GSMM is required: the *i*JO1366 GSMM for *E. coli* and *i*MM904 [15] for *S. cerevisiae* were used. Therefore, a total of 8 result sets were generated for two algorithms, two case studies and two search spaces (databases).

### 4.2 Data Preprocessing

Before running the algorithms, several pre-processing tasks needed to be performed. The first was to select and define the constraints of the problem, selecting the search space $\Sigma$, the initial set $S$ and the target compounds $T$. For both case studies, the target set is a singleton containing only the compound of interest, 1-butanol in the first case and vanillin in the second. For the substrate set, all metabolites included in the GSMMs were selected. This later will allow to integrate the obtained solutions with these models and evaluate their performance. The BiGG database [19] aided in the transformation of the species identifiers of the model to those in the databases. The species that did not match any cross-referencing were discarded.

The reference pathway of the 1-butanol synthesis was mostly present in the *i*JO1366 GSMM. So, to obtain alternative pathways we removed the following species: M_btcoa_c (Butanyl-CoA), M_btal_c (Butanal), M_b2coa_c (Crotonyl-CoA), M_3hbcoa_c (3-hydroxybuty), M_aacoa_c (Acetoacetyl-Coa). Additionally, every reaction connected to these compounds was also removed. The impact in the biomass value calculated using the FBA was minimal reducing to 0.977 (from 0.986). Removing these species will allow to find alternative paths from other internal metabolites of *i*JO1366 to 1-butanol, since an alternative solution to the identified in [1] is desired which may or may not be optimal against existing pathway. Note that the algorithms do not generate solutions including reactions to produce the initial substrate set since these are defined as supplied. Regarding the other case study, no modifications were made in the *i*MM904 GSMM.

A minor modification was done in the MetaCyc database, since it contains reactions with the metabolite pairs `NAD-P-OR-NOP/NADH-P-OR-NOP` which are instances of either NAD/NADH or NADP/NADHP. These reactions were unfolded to their correct instances. This is essential to infer the 1-butanol pathway, as several reactions of this pathway were expressed in this format. The KEGG Ligand database did not require pre-processing.

### 4.3 Algorithm Setup

Because of the combinatorial explosion of possible pathways, it is impossible to obtain every solution existing in a database size network using any of the algorithms. To compare the algorithms, the search space was split into subsets by *radius*. The *radius* is an integer that defines the minimum number of links (i.e., reactions) required to reach that reaction from an initial set of metabolites. This implies that a reaction that belongs to radius $i$ also belongs to $i + 1$, and therefore a sub-network $\Sigma_i$ of *radius* $i$ always complies to $\Sigma_i \subseteq \Sigma_{i+1}$.

With these reduced search spaces, solutions were computed using each of the algorithms. An attempt was made to obtain the entire set of candidate solutions for each *radius*, until either the proces crashes due to lack of memory or exceeds computational time allotted ($>$ 24 hours). To validate the solutions, FBA was used to maximize the yield of the target product and validate its feasibility integrating the solution into the respective GSMM.

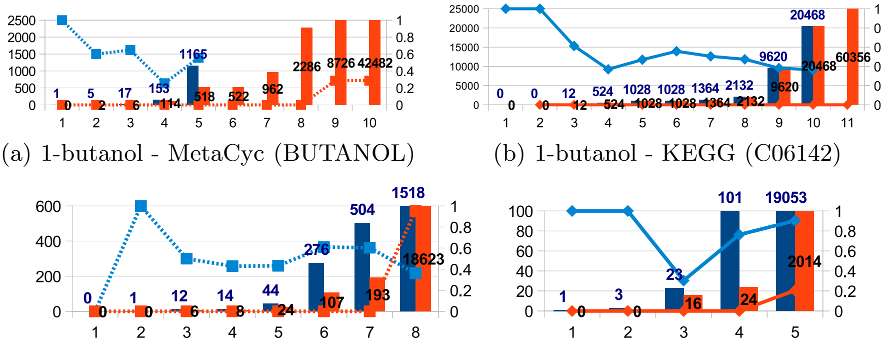
### 4.4 Results

Figure 1 shows the number of solutions computed and their feasibility. SSG is more limited than FP by the size of the search space. A major problem of the SSG algorithm is the high memory demand because of the power set computation. With the reduction of the power set size (only partial sets are computed), it still presents high memory demand to branch all the possible combinations. Moreover, the SSG computes every solution that satisfies Definition 4 which eventually leads to the computation of infeasible pathways.

Still, in general, the SSG shows better performance in the computation of solutions (Figure 2) mainly because of the branching technique which gives a major advantage to the computation time per solution because of the backtracking. As the algorithm moves to a candidate solution, the next solution reuses the previous partial solution. This results in a neglectable impact on the computation time per solution as the search space increases (i.e., increasing size of the radius). However, since the number of solutions exponentially grows with the increasing size of the search space, the total computation time increases.

The FP is capable to compute larger search spaces, being the major bottleneck the computation time per solution, since the internal Minimize routine has quadratic complexity to the number of reactions [2]. A scenario was also found where FP computes multiple distinct redundant solutions.

For every solution that satisfies the feasibility test, its performance was evaluated by integrating into the corresponding GSMM. The farthest *radius* that either algorithm was able to compute was selected for this process. For the



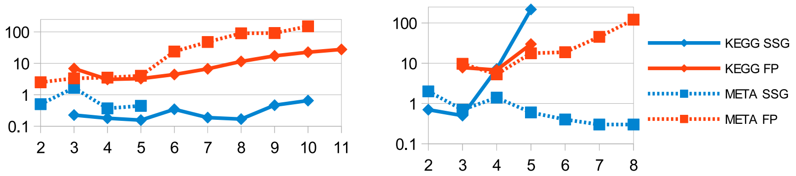
(a) 1-butanol - MetaCyc (BUTANOL)

(b) 1-butanol - KEGG (C06142)

(c) vanillin - MetaCyc (VANILLIN)

(d) vanillin - KEGG (C00755)

**Fig. 1.** Pathways computed for each of the problems by radius. The number of solutions on the left. The percentage of infeasible or redundant solutions on the right. Blue - SSG. Orange - FP.

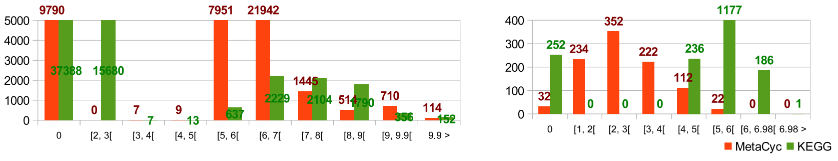


(a) 1-butanol

(b) vanillin

**Fig. 2.** Time cost (milliseconds) per each solution

1-butanol case, from the 42482 and 60356 solutions obtained from the FP algorithm, a total of 32692 and 22968 were compatible with the *i*JO1366 GSMM for search spaces of MetaCyc and KEGG, respectively. In the vanillin case, 944 of 974 computed solutions are valid (MetaCyc), being the numbers for KEGG of 1600 out of 1852. The 1-butanol case shown a massive amount of solutions mostly because of the NAD/NADH alternatives for many reactions.



(a) 1-butanol

(b) vanillin

**Fig. 3.** Histogram of theoretical yield values of 1-butanol in *i*JO1366 and vanillin in *i*MM904 . On the y-axis - number of solutions, x-axis yield range. Last value is the optimal solution (for better yield).

KEGG provided the solutions with highest yield, with 6.98 of vanillin in *i*MM904 , and 152 pathways with 9.99 for 1-butanol compared to 114 pathways from MetaCyc. There is a noticeable difference in the stoichiometry of the reactions between KEGG and MetaCyc in the 1-butanol pathways. A detailed view of the pathways obtained in this case study can be found in the supplementary material (http://darwin.di.uminho.pt/pacbb14-liu).

## 5 Conclusions and Future Perspectives

The algorithms analyzed both present errors in the computation of heterologous pathways. Although topologically they are correct, as they have the common goal which is to infer heterologous pathways (subnetworks) that satisfy the rules of initial substrates and target product, in a steady state point of view several examples may be infeasible. However, by using post-processing methods such as FBA, the correct solutions can be identified, which allows to correctly enumerate multiple pathways. The case study of 1-butanol shows that there are many viable routes for 1-butanol production in *i*JO1366 all with the same optimal yield. Moreover, even if a problem contains only a single optimal solution (e.g., vanillin in *i*MM904 ), examples of sub-optimal pathways also show a broad range of yield values many near the optimal. Other methods hardly can achieve such a range of feasible steady state heterologous pathways.

Thus, it is shown that although neither of the algorithms is readily suitable to compute steady state heterologous pathways for large databases, they are still able extract potential pathways, after targeted improvements in scalability. Additionally, they offer a generic method to infer pathways for multiple purposes, since they to not follow any strict objective function (e.g., yield or size).

**Acknowledgments.** The work is partially funded by ERDF - European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT (Portuguese Foundation for Science and Technology) within projects ref. COMPETE FCOMP-01-0124-FEDER-015079 and Strategic Project PEst-OE/EQB/LA0023/2013, and also by Project 23060, PEM - Technological Support Platform for Metabolic Engineering, co-funded by FEDER through Portuguese QREN under the scope of the Technological Research and Development Incentive system, North Operational.

## References

1. Atsumi, S., Cann, A.F., Connor, M.R., Shen, C.R., Smith, K.M., Brynildsen, M.P., Chou, K.J.Y., Hanai, T., Liao, J.C.: Metabolic engineering of Escherichia coli for 1-butanol production. Metabolic Engineering 10(6), 305–311 (2008)
2. Carbonell, P., Fichera, D., Pandit, S.B., Faulon, J.-L.: Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. BMC Systems Biology 6(1), 10 (2012)

3. Chatsurachai, S., Furusawa, C., Shimizu, H.: An in silico platform for the design of heterologous pathways in nonnative metabolite production. BMC Bioinformatics 13(1), 93 (2012)
4. Croes, D., Couche, F., Wodak, S.J., van Helden, J.: Metabolic PathFinding: inferring relevant pathways in biochemical networks. Nucleic Acids Research 33(Web Server issue), W326–W330 (2005)
5. Curran, K.A., Alper, H.S.: Expanding the chemical palate of cells by combining systems biology and metabolic engineering. Metabolic Engineering 14(4), 289–297 (2012)
6. de Figueiredo, L.F., Podhorski, A., Rubio, A., Kaleta, C., Beasley, J.E., Schuster, S., Planes, F.J.: Computing the shortest elementary flux modes in genome-scale metabolic networks. Bioinformatics 25(23), 3158–3165 (2009)
7. Faust, K., Croes, D., van Helden, J.: Metabolic pathfinding using RPAIR annotation. Journal of Molecular Biology 388(2), 390–414 (2009)
8. Faust, K., Dupont, P., Callut, J., Helden, J.V.: Pathway discovery in metabolic networks by subgraph extraction. Bioinformatics 26(9), 1211–1218 (2010)
9. Friedler, F., Tarján, K., Huang, Y., Fan, L.: Graph-theoretic approach to process synthesis: axioms and theorems. Chemical Engineering Science 47(8), 1973–1988 (1992)
10. Friedler, F., Tarjan, K., Huang, Y.W., Fan, L.: Graph-theoretic approach to process synthesis: Polynomial algorithm for maximal structure generation. Computers & Chemical Engineering 17(9), 929–942 (1993)
11. Friedler, F., Varga, J., Fan, L.: Decision-mapping: A tool for consistent and complete decisions in process synthesis. Chemical Engineering Science 50(11), 1755–1768 (1995)
12. Hansen, E.H., Møller, B.L., Kock, G.R., Bünner, C.M., Kristensen, C., Jensen, O.R., Okkels, F.T., Olsen, C.E., Motawia, M.S., Hansen, J.R.: De Novo Biosynthesis of Vanillin in Fission Yeast (Schizosaccharomyces pombe) and Baker's Yeast (Saccharomyces cerevisiae). Applied and Environmental Microbiology 75(9), 2765–2774 (2009)
13. Lee, D.-Y., Fan, L.T., Park, S., Lee, S.Y., Shafie, S., Bertók, B., Friedler, F.: Complementary identification of multiple flux distributions and multiple metabolic pathways. Metabolic Engineering 7(3), 182–200 (2005)
14. Machado, D., Soons, Z., Patil, K.R., Ferreira, E.C., Rocha, I.: Random sampling of elementary flux modes in large-scale metabolic networks. Bioinformatics (Oxford, England) 28(18), i515–i521 (2012)
15. Mo, M.L., Palsson, B.O., Herrgård, M.J.: Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC Systems Biology 3, 37 (2009)
16. Orth, J.D., Thiele, I., Palsson, B.O.: What is flux balance analysis? Nature Biotechnology 28(3), 245–248 (2010)
17. Pharkya, P., Burgard, A.P., Maranas, C.D.: OptStrain: A computational framework for redesign of microbial production systems. Genome Research 814, 2367–2376 (2004)
18. Rahman, S.A., Advani, P., Schunk, R., Schrader, R., Schomburg, D.: Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics (Oxford, England) 21(7), 1189–1193 (2005)
19. Schellenberger, J., Park, J.O., Conrad, T.M., Palsson, B.O.: BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11, 213 (2010)