Flux Measurement Selection in Metabolic Networks

Wout Megchelenbrink, Martijn Huynen, and Elena Marchiori

Radboud University, Nijmegen,
The Netherlands

Abstract. Genome-scale metabolic networks can be reconstructed using a constraint-based modeling approach. The stoichiometry of the network and the physiochemical laws still enable organisms to achieve certain objectives -such as biomass composition- through many various pathways. This means that the system is underdetermined and many alternative solutions exist. A known method used to reduce the number of alternative pathways is Flux Balance Analysis (FBA), which tries to optimize a given biological objective function. FBA does not always find a correct solution and for many networks the biological objective function is simply unknown. This leaves researchers no other choice than to measure certain fluxes. In this article we propose a method that combines a sampling approach with a greedy algorithm for finding a subset of k fluxes that, if measured, are expected to reduce as much as possible the solution space towards the 'true' flux distribution. The parameter k is given by the user. Application of the proposed method to a toy example and two real-life metabolic networks indicate its effectiveness. The method achieves significantly more reduction of the solution space than when k fluxes are selected either at random or by a faster simple heuristic procedure. It can be used for guiding the biologists to perform experimental analysis of metabolic networks.

1 Introduction

An important goal for researchers in systems biology is to understand the properties of metabolic networks. These networks are built in a bottom-up fashion using various biological sources, such as genome annotations, metabolic databases and biochemical information [19]. Most metabolic networks are large and complex, having many alternative pathways for constructing certain metabolites. Therefore they are often modeled in a constraint-based fashion [24,14,13]. This approach enables researchers to perform quantitative analysis within a validated mathematical model. For instance, COBRA [21] is a computational toolbox based on this approach, that enables researchers to model and infer metabolic networks. Metabolic networks can be defined by two types of constraints. The first is the so-called mass-balance constraint in (1).

$$\frac{dx}{dt} = Sv, \tag{1}$$

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 214–224, 2011.

where dx/dt is the change in metabolite concentration over time. S is the stoichiometry matrix, and v is the flux vector. Assuming a steady-state, (1) simplifies to $Sv = 0$. The thermodynamic constraint (2 is the second type, which limits the upper and lower bounds of the flux rates.

$$v_{j,min} \leq v_j \leq v_{j,max}, \forall j \in J, \tag{2}$$

where v_j is the flux value of reaction j, and J denotes the set of reactions. Most of the networks contain more reactions than metabolites, leaving the system underdetermined and resulting in many possible solutions, that is, alternative flux distributions. In order to find the 'true' flux distribution under certain growth conditions, researchers measure certain fluxes to tighten the constraints; thereby reducing the solution space towards the 'true' flux distribution. One of the challenges we address in this paper is to detect a small set of k reactions that, when measured, will maximally reduce the solution space. Here k is a parameter selected by the user.

2 Related Work

Two early papers about discovering optimal flux measurements [18,17] suggest measuring those fluxes that are least sensitive to experimental error. The authors show that an upper bound for this sensitivity can be approximated solely on stoichiometry. When actual information is known on measurement errors, the sensitivity can be computed more accurately. However, these measurement are not optimized for reducing the solution space.

More recent methods for determining metabolic fluxes often optimize a biological objective function such a growth or ATP production. A well-known method that uses this strategy is flux balance analysis [9,12,23]. However for eukaryotic cells a biological objective function is often not easy to determine. An alternative approach often used in perturbed networks is the minimization of metabolic adjustment [22], involving the minimization of the Euclidean distance between the wild-type and perturbed network.

In the setting considered in this paper we assume that no information other than the mass-balance and thermodynamic constraints, is available. In particular, we do not use any biological objective function or external sources of knowledge such as gene expression. The goal is to find a set of k fluxes that, if measured, will reduce as much as possible the search space obtained from constraints (1) and (2). To the best of our knowledge, this is the first time such a problem is tackled in the context of metabolic networks analysis. However there are methods that try to tackle the related problem of finding the shape and size of the solution space by randomized sampling [10,20]. We will show that these type of methods are very useful also for addressing the problem we want to solve.

In [25] the authors use a sampling approach to find the size and shape of the steady-state solution space. The authors demonstrate which reactions are sensitive to the V_{max} conditions using a singular value decomposition on the human red blood cell network [8]. The authors of [15] show that the sampling

approach can be used to find probability distributions for all fluxes, that it can be used to measure pairwise correlation coefficients and to compute the network wide effects of changes in flux variables. Braunstein et al. [3] show an alternative approach to approximate the volume and shape of the convex polytope using a message-passing algorithm based on belief propagation. Almaas et al. [1] have used random sampling on the E. coli network to show that there is a core set of reactions carrying a high flux, termed the 'high-flux backbone'. Finally flux variability analysis (FVA) [7,11] has been used to explore alternative optima and network redundancy.

3 Research Problem

The most reliable method to reduce the solution space is to perform experimental measurements to find the real flux values. The goal of this paper is to help the biologists to decide which k reactions to measure in order to get as close as possible to the 'true' flux distribution the organism uses, where k is a user given parameter.

The solution space of the system of equations in (1) and inequalities in (2) forms a bounded convex polytope. The number of alternative flux distributions can be expressed as the hypervolume of this polytope. Computing the exact volume of a convex polytope is a NP-hard problem and has been shown to be infeasible for dimensions bigger than 10 [2]. Since we don't need the exact volume but the smallest volume resulting when measuring a flux, we only need to estimate a relative volume. The authors of [25] have shown that the relative volume of a more constrained polytope P_c relative to the original model P_o can be approximated as follows:

$$\hat{V}_{rel} = \frac{|P_c|}{|P_o|}, \tag{3}$$

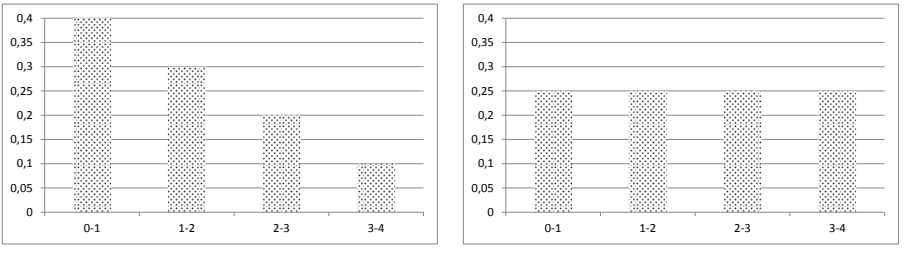
where $|P_o|$ is the number of sample points taken from P_o and $|P_c|$ is the number of these sample points that are also in the more constrained model P_c. Whenever a flux is measured, the initial bounds in the constrained model of the metabolic network become those of the measured value $\pm$ a term quantifying the uncertainty of the measurement. As a consequence, also the hypervolume of the polytope is reduced. The problem is that we don't know the value resulting from such experimental measurement. Therefore we resort to a sampling histogram of the network for a given flux: the number of samples within a certain bin reflects the expected volume of the polytope when the flux is constrained to assume its value in that particular bin. Formally, the expected volume when flux j is selected for being measured is

$$\hat{V}_j = \sum_{b=1}^{n} V_{j(b)}^2, \tag{4}$$

where n is the number of bins considered, $V_{j(b)}$ is the relative volume for bin b in flux j and $\sum_{b=1}^{n} V_{j(b)} = 1, \forall j$. This expected volume also reflects the probability

that a sample is in that bin. Thus a flux having similar values across all bins will, when measured, yield the highest expected volume reduction. An example illustrates the observations above.

Example 1. Consider the histograms in Fig. 1. Figure 1a represents a flux yielding a low expected volume. Specifically, if we measure this flux, the expected volume of the polytope is 0.3. A flux yielding higher expected volume reduction when measured is shown in Fig. 1b: indeed, for that flux, $\hat{V}$ is minimal, equal to 0.25, which corresponds to the volume of any bin.



(a) Flux with $\hat{V} = 0.30$ (b) Flux with $\hat{V} = 0.25$

Fig. 1. Finding the flux with minimum expected volume

We want to find a set K^* consisting of k fluxes, that minimizes the expected volume, that is, we want to find the optimum of the following objective function:

$$\arg\min_{K \subseteq J} \hat{V}(K), \tag{5}$$

where J denotes the set of all fluxes in the considered metabolic network, K denotes a subset J of size k, and $\hat{V}(K)$ denotes the expected volume resulting from the measurements of the reactions in the set K. A greedy method for tackling this optimization problem is proposed in the next section.

4 Methods

For $k = 1$ one can tackle the optimization problem (5) using the formula (4) for finding the flux with minimum expected volume. However, using laboratory techniques such as isotope labeling enables researchers to measure multiple fluxes at once [16]. Therefore it is practically more useful to select $k > 1$ fluxes. An exact approach to solve optimization problem (5) amounts to search among all possible k flux combinations the one that minimizes $\hat{V}$. This exhaustive search becomes intractable on any genome-scale metabolic network even for low values of k, due to the combinatorial explosion of the number of subsets of reactions.

A fast heuristic method would be to select the k fluxes having smallest expected volume, as computed using equation (4). We call this method Reaction Minimizing Expected Volume Naive (RMEV-N in short). As shown in the computational analysis provided in the next section, this approach yields results of low quality. This is due to the fact that the shape and volume of the polytope changes whenever a flux is measured while all the computed predictions use the same initial model of the considered metabolic network. Therefore we consider a more involved, greedy search method, called Reaction Minimizing Expected Volume Greedy (RMEV-G in short) which is described in the next section.

4.1 RMEV-G: Reaction Minimizing Expected Volume Greedy

RMEV-G consists of three main phases. First, it reduces the ranges of the fluxes using a method for constraint domain reduction implemented in the COBRA toolbox. The resulting ranges will be used for computing a set of n bins $b_1 \dots b_n$ for the range of each flux. Next, it generates a set of points from the solution space of the mass-balance (see (1)) and thermodynamic (see (2)) constraints, by means of a uniform sampling procedure. Specifically, we use an implementation of the ACHR sampler algorithm [10,21] to generate such a set of sampled solutions. This set will be used for computing the expected volume of a flux, using equation (4). Finally, RMEV-G uses the resulting ranges and sampled solutions for selecting k fluxes $j_1, \dots, j_k$ using the following greedy search procedure.

- j_1. The reaction with lowest $\hat{V}$ is picked as the first selected reaction j_1. This can be done by computing the expected volume of each flux using equation (4) and selecting the flux with minimum expected volume.
- j_i, *with* $i > 1$. At iteration i, the method tries to find the reaction j_i that minimizes the expected volume $\hat{V}$ given that the reactions $j_1 \dots j_{(i-1)}$ have been selected. Since we only know that $j_1, \dots, j_{i-1}$ have been selected, but we do not know in which bins the selected fluxes will occur, we have to consider each bin as a possibility. Therefore we construct a search tree as follows.

Search Tree Construction. Nodes of this tree are (labeled with) fluxes and edges are (labeled with) bins. Each edge has a weight, equal to the probability of the parent node (flux) to be in the bin corresponding to that edge given the bins that have been selected in the path from the root to that edge (for instance, the probability that the flux in Fig. 1a is in bin [1,2] is 0.3).

1. The root of this tree is j_1.
2. Each node has n children, one for each bin (edge). Each of these children is the flux having minimum expected volume in the reduced solution space resulting from the selection of fluxes and bins along the path from the root to that node. The expected volume of that flux in this reduced space can be computed using the subset of the sampled solutions obtained by discarding those that are not consistent with the new restricted ranges of the fluxes occurring in that path.

An example of such a search tree in given Fig. 2, with 3 bins and $k = 3$ fluxes to be selected. The search tree is used to select j_i by means of a weighted majority vote criterion applied to the nodes occurring at depth i.

Majority Vote Selection Criterion. For each flux, the weighted sum of the occurrences of that flux at depth i in the search tree is computed, where each occurrence is weighted by the weight of the edge linking that occurrence with its parent. Then the flux with maximum weighted sum is selected (ties are broken randomly). The output of RMEV-G is an ordered list of k fluxes and a resulting expected volume after measuring each flux given the previous ones that have been selected. The computational complexity of RMEV-G is $O(mn^k)$, where m is the total number of fluxes, n the number of bins, and k the number of fluxes to be selected. Therefore the method is in general applicable for small values of k and n.

Example 2. Consider the example search tree in Fig. 2, with 3 bins and $k = 3$ fluxes to be selected. The first reaction selected is f_4, having minimum $\hat{V}$. The three bins of f_4 are split and the reaction having minimum $\hat{V}$ is multiplied by its prior probability. The reaction with highest weighted vote is used in all subsequent iterations.

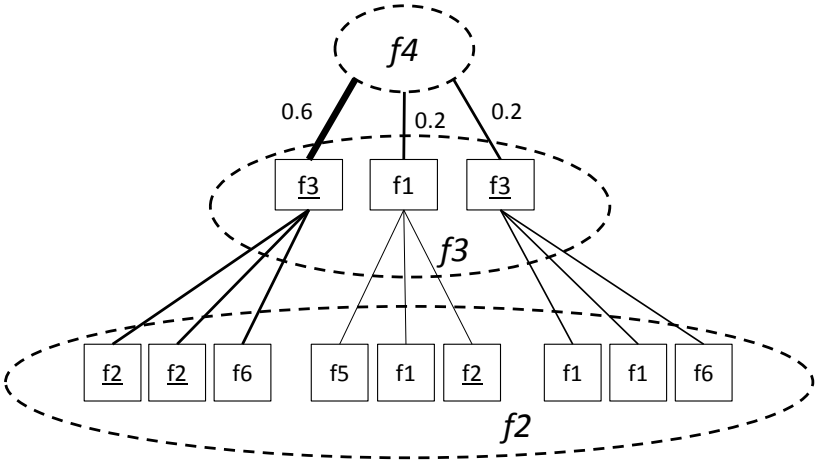


Fig. 2. Schematic representation of a search tree with three bins and three fluxes. The line width is proportional to the probability mass in each bin and thus to the weight of each vote. Underlined fluxes are those selected using weighted majority vote. In particular, in the second iteration, f_3 is chosen because it has the maximum weighted vote, equal to $0.6 + 0.2$.

In the next section we test the effectiveness of RMEV-G on artificial and real-life metabolic networks, and measure the reduced volume versus that achieved by applying faster algorithms like RMEV-N and random selection.

5 Experimental Analysis

In order to test comparatively the effectiveness of the proposed method we have conducted computational experiments on the following three networks.

1. Toy model. This is a small network explained in [25]. It contains only 8 reactions and 5 metabolites.
2. Red blood cell (RBC). This metabolic network is a constraint-based network based on the kinetic model from [8]. We have used the model that was available as supplementary material from [3].
3. E. coli central metabolism (E. coli c.m.). This is a condensed version of the genome-scale E. coli network. It contains reactions and metabolites from central metabolism and is used also in [13]. It can be publicly downloaded[1].

We considered the following three methods.

- **RMEV-G.** The proposed greedy method previously described.
- **Reaction Minimizing Expected Volume Random (RMEV-R).** This algorithm selects fluxes randomly, adding the extra constraint that fluxes can not be fully coupled [4,6]. Fully coupled reactions often occur in a linear path and their fluxes are by definition the same. Therefore it is useless to measure multiple reactions in such a set. After selecting k fluxes not containing fully coupled reactions, we used the same bins as for the greedy approach and traverse the search tree in the exact same way. At every level of the tree, from $i = 1, \ldots, i = k$ we computed the expected volume using 3.
- **Reaction Minimizing Expected Volume Naive (RMEV-N).** This algorithm amounts to performing only the first iteration of RMEV-G and then choosing the top k fluxes having smallest expected volume and not containing fully coupled reactions. Note that this method measures the expected volume for all fluxes based on the original solution space. Therefore it does not take into account the changed shape and size of the solution space after selection of a flux.

We applied these methods to each of the considered networks, and compared the expected volumes resulting from the selection of $k \leq 5$ fluxes using $n = 5$ bins. For each network, $N = 1000$ runs of RMEV-R were performed. Results of these runs were used to compute an empirical p-value, as the fraction of runs were RMEV-R achieved smaller expected volume than RMEV-G.

$$\sum_{i=1}^{N} \frac{I(\hat{V}_1 > \hat{V}_2)}{N}, \tag{6}$$

where I denotes the indicator function, 1 denote method RMEV-G and 2 method RMEV-R.

[1] `http://www-bioeng.ucsd.edu/research/research_groups/gcrg/organisms/ecoli/ecoli_sbml.html`

5.1 Results

Table 1 reports the results of our experiments. They show that RMEV-G performs better than the other methods. In particular, results of the experiments indicate that RMEV-N performs worse than RMEV-G. This is expected since the latter algorithm considers a reduced solution space after each selection of a flux.

Table 1. Results on the toy model and two real-life metabolic networks. Column "Model" contains the considered metabolic networks, "K" depicts the number of fluxes determined, "Size" the network size (metabolites, reactions), RMEV-R, RMEV-N, and RMEV-G the expected volume computed using the three methods, where for RMEV-R, the average over the considered runs is reported. Finally, "P-value" denotes the empirical p-value between RMEV-G and RMEV-R.

Model	Size	K	RMEV-R	RMEV-N	RMEV-G	P-value
Toy model	5, 8					
		1	0.2402	0.2363	0.2363	0
		3	0.0194	0.0193	0.0193	0
		5	0.0069	0.0069	0.0069	0.6450
RBC	34, 46					
		1	0.4156	0.2104	0.2104	0
		3	0.1047	0.0666	0.0229	0.0020
		5	0.0356	0.0079	0.0044	0
E. coli c.m.	62,75					
		1	0.7127	0.2001	0.2001	0
		3	0.3794	0.0702	0.0334	0
		5	0.1704	0.0325	0.0083	0.0630

The last column in the table contains the p-value computed using (6). The relative high p-value for the toy model and $k = 5$ can be justified by the small dimension of the solution space (which is 3), meaning that there is little room for more reduction of the solution space. On the two real-life metabolic networks small p-values are obtained, except for the E. coli network, where a p-value of 0.0630 is computed. This could be possibly due to the constraint that the selected fluxes should not contain fully coupled reactions. This remains to be investigated in more depth. Figure 3 shows the performance of the methods on the largest network considered, that is, the E. coli one. The theoretical lower bound for $\hat{V}$ is also plotted, computed as $lb_k = 1/n^k$. The plots clearly show that RMEV-G has a very good performance, with expected volume closer to the lower bound than the other two methods.

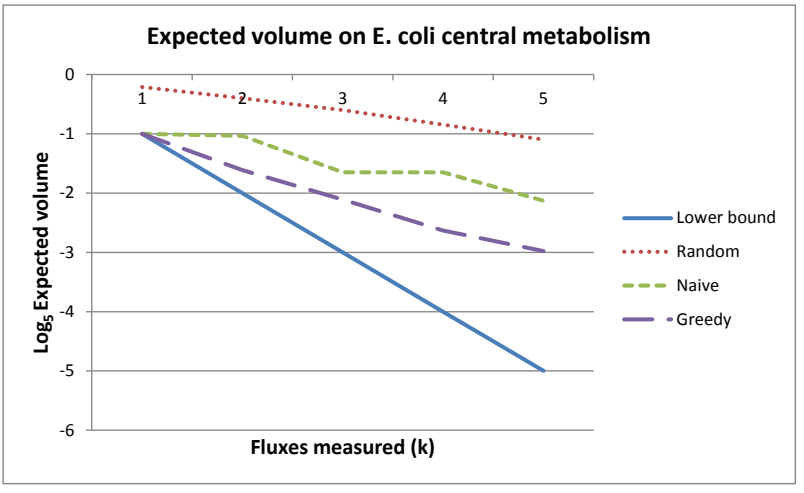


Fig. 3. Performance of the three methods and the theoretical lower bound of $\hat{V}$ on the E. coli central metabolism network. The x-axis reports the number of fluxes selected, the y-axis the $\log_5$ of the expected volume, where the base 5 is the number of bins.

6 Conclusion and Discussion

The experimental analysis shows that RMEV-G can be used to help selecting fluxes for measurement in the laboratory. In all networks examined RMEV-G performs better than the two other methods discussed, because it considers the new solution space after selecting each subsequent flux. In particular, RMEV-G reduces the solution space much more compared to random selection, even when the latter is restricted not to contain fully coupled reactions. Note that by the definition of RMEV-G it takes reaction coupling implicitly into account, because fully coupled fluxes will have sample points that coincide and therefore will give a higher expected volume than any other flux. An important point to note is that RMEV-G is based on only stoichiometry and flux bounds and is therefore highly sensitive to how the initial network is constrained. In the future, we will try to incorporate biological data, such as gene expression in order to alleviate this problem and to make predictions more accurate.

7 Future Work

Although the results of the preliminary experiments conducted in this work indicate that the proposed greedy method is effective, we are going to extend the experimental analysis by considering genome-scale metabolic networks. Moreover, we want to test the method on a measured network, to see if our predictions agree with the measured values.

Furthermore, we are currently working on an extension of the method that incorporates prior information from gene expression data. Another interesting possibility we are currently investigating is combining our algorithm with an approach that tries to minimize error amplification, like for instance the one introduced in [18,17].

Another issue we want to address is the time complexity of the method. The exponential nature (on the number of bins and number of selected fluxes) poses a problem if one wants to measure a larger number of fluxes. For instance, a possible solution could be to parallelize the algorithm by running in parallel the generation of (groups of) children nodes at each depth of the search tree prior to the integration of the decisions computed from the nodes using the majority vote criterion.

Finally, we want to investigate the relation between our method and the Bayesian experimental design approach [5].

Acknowledgments. The authors would like to thank Richard Notebaart, Sergio Rossell and Radek Szklarczyk for their valuable comments. This work has been funded by the Netherlands organization for scientific research (NWO) and the Netherlands organization for health research and innovation in healthcare (ZonMW).

References

1. Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z.N., Barabási, A.-L.: Global organization of metabolic fluxes in the bacterium escherichia coli. Nature 427(6977), 839–843 (2004)
2. Beeler, B., Enge, A., Fukuda, K., Lthi, H.-J.: Exact volume computation for polytopes: a practical study. In: 12th European Workshop on Computational Geometry, Muenster, Germany (1996)
3. Braunstein, A., Mulet, R., Pagnani, A.: Estimating the size of the solution space of metabolic networks. BMC Bioinformatics 9(1), 240 (2008)
4. Burgard, A.P., Nikolaev, E.V., Schilling, C.H., Maranas, C.D.: Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Research 14(2), 301–312 (2004)
5. Chaloner, K., Verdinelli, I.: Bayesian experimental design: A review. Statistical Science 10(3), 273–304 (1995)
6. David, L., Marashi, S.-A., Larhlimi, A., Mieth, B., Bockmayr, A.: FFCA: a feasibility-based method for flux coupling analysis of metabolic networks. BMC Bioinformatics 12(1), 236 (2011)
7. Gudmundsson, S., Thiele, I.: Computationally efficient flux variability analysis. BMC Bioinformatics 11, 489 (2010)
8. Jamshidi, N., Edwards, J.S., Fahland, T., Church, G.M., Palsson, B.O.: Dynamic simulation of the human red blood cell metabolic network. Bioinformatics 17(3), 286–287 (2001)
9. Kauffman, K.J., Prakash, P., Edwards, J.S.: Advances in flux balance analysis. Current Opinion in Biotechnology 14(5), 491–496 (2003)

10. Kaufman, D.E., Smith, R.L.: Direction choice for accelerated convergence in hit-and-run sampling. Operations Research 46(1), 84–95 (1998)
11. Mahadevan, R., Schilling, C.H.: The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering 5(4), 264–276 (2003)
12. Orth, J.D., Thiele, I., Palsson, B.O.: What is flux balance analysis? Nature Biotechnology 28(3), 245–248 (2010)
13. Palsson, B.O.: Systems Biology: Properties of Reconstructed Networks, 1st edn. Cambridge University Press (2006)
14. Price, N.D., Reed, J.L., Palsson, B.O.: Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature Reviews Microbiology 2(11), 886–897 (2004)
15. Price, N.D., Schellenberger, J., Palsson, B.O.: Uniform sampling of Steady-State flux spaces: Means to design experiments and to interpret enzymopathies. Biophysical Journal 87(4), 2172–2186 (2004)
16. Sauer, U.: Metabolic networks in motion: 13C-based flux analysis. Molecular Systems Biology 2, 62 (2006)
17. Savinell, J.M., Palsson, B.O.: Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. Journal of Theoretical Biology 155(2), 201–214 (1992)
18. Savinell, J.M., Palsson, B.O.: Optimal selection of metabolic fluxes for in vivo measurement. II. Application to escherichia coli and hybridoma cell metabolism. Journal of Theoretical Biology 155(2), 215–242 (1992)
19. Schellenberger, J., Lewis, N.E., Palsson, B.O.: Elimination of thermodynamically infeasible loops in steady-state metabolic models. Biophysical Journal 100(3), 544–553 (2011)
20. Schellenberger, J., Palsson, B.O.: Use of randomized sampling for analysis of metabolic networks. The Journal of Biological Chemistry 284(9), 5457–5461 (2009)
21. Schellenberger, J., et al.: Quantitative prediction of cellular metabolism with constraint-basded models: the cobra toolbox v2.0. Nature Protocols 6(9), 1290–1307 (2011)
22. Segré, D., Vitkup, D., Church, G.M.: Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences of the United States of America 99(23), 15112–15117 (2002)
23. Smallbone, K., Simeonidis, E.: Flux balance analysis: a geometric perspective. Journal of Theoretical Biology 258(2), 311–315 (2009)
24. Varma, A., Palsson, B.O.: Metabolic flux balancing: Basic concepts, scientific and practical use. Nature Biotechnology 12, 994 (1994)
25. Wiback, S.J., Famili, I., Harvey, J., Greenberg, H.J., Palsson, B.O.: Monte carlo sampling can be used to determine the size and shape of the steady-state flux space. Journal of Theoretical Biology 228(4), 437–447 (2004)