Vol.2021-BIO-67 No.2
2021/9/30

Dynamic solution space division-based methods for calculating reaction deletion strategies for constraint-based metabolic networks for substance production: DynCubeProd

Yier Ma[1,a)] Takeyuki Tamura[1,b)]

Abstract: Flux balance analysis (FBA) is a crucial method to analyze large-scale constraint-based metabolic networks and computing design strategies for strain production in metabolic engineering. However, as it is often non-straightforward to obtain such design strategies to produce valuable metabolites, many tools have been proposed based on FBA. Among them, GridProd, which divides the solution space into small squares by focusing on the cell growth rate and the target metabolite production rate to efficiently find the reaction deletion strategies, was extended to CubeProd, which divides the solution space into small cubes. However, as GridProd and CubeProd naively divide the solution space into equal sizes, even places where solutions are unlikely to exist are examined. To address this issue, we introduce dynamic solution space division methods based on CubeProd for faster computing by avoiding searching in places where the solutions do not exist. We applied the proposed method DynCubeProd to iJO1366, which is a genome-scale constraint-based model of *Escherichia coli*. Compared with CubeProd, DynCubeProd significantly accelerated the calculation of the reaction deletion strategy for each target metabolite production. In addition, under the anaerobic condition of iJO1366, DynCubeProd could obtain the reaction deletion strategies for almost 40% of the target metabolites that the elementary flux vector-based method, which is one of the most effective methods in existence, could not. This study was published in https://doi.org/10.3389/fbinf.2021.716112 [4].

Keywords: metabolic network, flux balance analysis, constraint-based model, linear programming, algorithm

1. Introduction

Metabolic engineering is a DNA recombination-based technology proposed in 1991 to improve the designated substance production and the cell properties by manipulating and introducing specific biochemical reactions [1], [7]. In many cases, current metabolic engineering technology focuses on the utilization of microorganisms. In metabolic engineering analysis, metabolic pathways in organisms are often represented by metabolic networks, in which nodes represent metabolite molecules and biochemical reactions. Any two metabolites (biochemical reactions) cannot be directly connected, and a metabolite must be connected to at least two biochemical reactions. The biochemical reactions can be irreversible or reversible. Nodes of external metabolites form the input and output nodes of the entire network.

Constraint-based modeling is a mathematical method to identify the best solution within a set of possible choices subject to pre-specified constraints [5]. Constraint-based modeling methods, such as linear programming (LP) and mixed integer linear programming, are widely used effective optimization techniques. Flux balance analysis (FBA) is one such widely used constraint-based modeling method with stoichiometric-based modeling of metabolism for the analysis of genome-scale metabolic models (GSMM) [5].

In the constraint-based models of metabolic networks, the cell growth reaction and the target metabolite produc-

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan
[a)] mayier@kuicr.kyoto-u.ac.jp
[b)] tamura@kuicr.kyoto-u.ac.jp

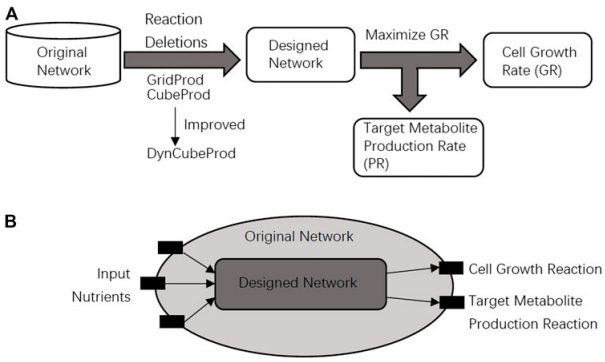


Fig. 1 **(A)** In the problem setting of this study, it is required that convert the original network is converted into the designed network that achieves growth coupling, where cell growth and target metabolite production are simultaneously achieved. **(B)** The designed network should be obtained from the original network through reaction deletions. The black blocks are nutrient uptake reactions, the cell growth reaction, or the target metabolite production reaction. The speed of the cell growth reaction and target metabolite production reaction are represented by GR and PR, respectively. Light gray and dark gray represent the designed network and the original network, respectively.

tion reactions are of particular interest. The cell growth reaction has been virtually designed to simulate the efficient conversion of uptake resources into cellular energy and chemical components, which support cell growth in response to selection pressure to construct the system in the most plausible physiological state [5]. The target metabolite production reaction produces a chemical of interest. We define growth rate (GR) as cell growth reaction speed and production rate (PR) as the target metabolite production reaction speed.

Growth coupling is a fundamental design principle in metabolic engineering and computational strain design. The purpose of growth coupling is to make the target metabolite a mandatory by-product of the cell growth reaction. We say that growth coupling is achieved if the target metabolite is produced when cell growth is maximized as shown in **Fig. 1 (A)**.

In this study, the core and basic problems are to find a growth coupling method for the target metabolite production through reaction deletions. The method should produce as much target metabolite as possible by modifying the metabolic network when GR is maximized as the objective function. The relationship between the original network and the designed network is shown in Fig. 1 (B). We can delete reactions by setting their speeds as zero in the network modification strategies. Based on the most basic problem above, the following sub-problems are derived. The first is to find knockout strategies for as many different target metabolites as possible. The second is to find knockout strategies for the networks under different input conditions such as aerobic or anaerobic conditions.

The most basic and pioneer algorithm for this purpose is OptKnock, which is a bilevel optimization-based method that identifies knockout strategies that result in the maximum PR when GR is maximized. The inner optimization performs the flux (reaction speed) allocation with regard to the optimization of cellular objectives (e.g., maximization of biomass yield and MOMA) [2]. The outer optimization maximizes the bioengineering objective (e.g., chemical production) [2]. However, because the computation time of OptKnock is proportional to an exponential function of the network size, in many cases, its computation is not completed within a realistic timeframe for GSMMs [10]. Therefore, many algorithms have been proposed to speed up the process for the efficient computation of the reaction deletion strategies.

Considering that finding the optimal strategy is NP-hard, it is reasonable to only find out the strategy that meets the expected requirements. For example, the elementary flux vector (EFV)-based method determines reaction deletion strategies in which cell growth forces the production of the target metabolite, and the success ratio of this method was very high under both anaerobic and aerobic conditions for several microbial models [11]. GridProd efficiently computes the design of minimum metabolic networks by using bilevel optimization approach with picking two-dimensional limits and gridding the constraint space [8]. CubeProd divides the entire constraint space into small cubes and gave good results for GSMM with extreme constraints (e.g., anaerobic condition) [9]. The EFV-based method, GridProd, and CubeProd enable the calculation of reaction deletion strategies for many target metabolites that cannot be calculated using the previously developed methods. However, for *Escherichia coli* under anaerobic conditions, the reaction deletion strategies could not be obtained for many target compounds. In particular, for GridProd and CubeProd, the bottleneck was the computing speed. Therefore, it was expected to extend GridProd or CubeProd to shorten the computation time.

In this study, we developed DynCubeProd that improves the computation speed of CubeProd. DynCubeProd employs a dynamic strategy for the cube sizes to obtain the same results as CubeProd; however its compu-

tation speed is much faster. The reaction deletion strategies obtained by DynCubeProd also supplement those of the EFV-based method under certain conditions. Under anaerobic conditions, we obtained the reaction deletion strategies for close to 40% of the target metabolites for which the EFV-based method could not determine strategies.

2. Materials and Methods

2.1 Problem definition

The general formalization of constraint-based modeling is as follows [5]:

minimize (or **maximize**):

$$f(x)$$

subject to:

$$h(x) = 0$$
$$g(x) \leq 0$$
$$x \in S$$

x is an n-dimensional variable. $f(x)$ is the objective function to minimize or maximize. S is the set from which the variable vector x ranges. $h(x)$ and $g(x)$ are the constraints that must be satisfied as equalities or one-side inequalities, respectively.

The general form of the FBA is as follow:

maximize

$$f(x)$$

subject to:

$$Sx = 0$$
$$LB \leq x \leq UB$$

$x \in R^n$ is an n-dimensional variable. $f(x)$ is the objective function, which in many cases is GR. $S \in R^{m \times n}$ is the stoichiometric matrix corresponding to m metabolites and n reactions in the constraint-based models. LB and UB impose the lower and upper bounds of each $x \in x$. For example, a flux for irreversible reactions x_i is constrained as $x_i \geq 0$.

Our goal was to find reaction deletion strategies for growth coupling of target metabolite production. Let $K = \{v_j | v_j \in V\}$ be a set of reactions to be knocked out, where V is a set of n reactions. Then, the definition of the main problem of this study arises.

Given

$$S, LB, UB, v_{growth}, v_{target}, x_{growth}^{min}, x_{target}^{threshold}$$

Find

$$K$$

such that

$$x_{growth} \geq x_{growth}^{min} \text{ and } x_{target} \geq x_{target}^{threshold}$$

maximize

$$f(x)\ (=x_{growth})$$

subject to:

$$Sx = 0$$
$$\{\ x = 0 \text{ if },\ x \in K$$

LB $\leq x \leq UB$, otherwise.

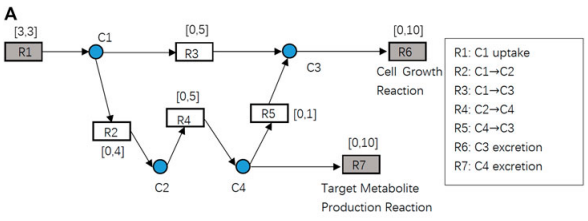
When $x_{growth} \geq x_{growth}^{min}$ and $x_{target} \geq x_{target}^{threshold}$ is satisfied, we consider K achieves growth coupling, where $GR = x_{growth}$ for $v_{growth} \in V$ and $PR = x_{target}$ for $v_{target} \in V$ hold.

2.2 Example for Problem Definition

A toy example of the constraint-based model with 11 nodes is shown in **Fig. 2 (A)** to illustrate the problem definition explained above. The rectangular nodes $\{R1, R2, \ldots, R7\}$ are chemical reactions. $R7$ is the target metabolite production reaction and $R6$ is the cell growth reaction. The substrates and products of the reactions are shown on the right side of Fig. 2 (A). The gray rectangular nodes are external reactions which play roles of input and output of the entire network and the white rectangular nodes are internal reactions, each of which connect at least two metabolite nodes with different directions. The intervals next to the rectangular nodes are the lower and upper bounds of reaction speeds. The circular nodes are internal metabolites that connect rectangular nodes.

Suppose that $v_{growth}^{min} = v_{target}^{threshold} = 1$ is given. When a reaction deletion strategy is given and $GR = x_6$ is maximized, if $GR \geq 1$ and $PR = x_7 \geq 1$ hold, then we consider that growth coupling is achieved. Because this example is very simple, such a reaction deletion strategy can be easily determined through brute force enumeration. Deleting $R3$ is the optimal solution for this toy example.

Fig. 2 (B) shows the results of each knockout strategy applied to this toy example. Because deleting R4 is practically equivalent to deleting R2, deleting R4 is omitted in the table. When none of the reactions are deleted, that is, $K = \phi$, GR=3 will be obtained but PR will be 0. When $K = \{R2\}$ or $K = \{R5\}$, the same result will be obtained. According to the definition of the problem above, such knockout strategies are not acceptable. When $K = \{R3, R5\}$, PR=3 is obtained, but GR will be 0. When $K = \{R3\}$, both $GR = 1 \geq 1$ and $PR = 2 \geq 1$ are obtained. Therefore, $K = \{R3\}$ is a feasible solution because it achieves growth coupling. However, such a brute force method cannot be applied to GSSMs owing



B

Deletion	R1	R2	R3	R4	R5	R6	R7
none	3	0	3	0	0	3	0
R3	3	3	0	3	1	1	2
R2	3	0	3	0	0	3	0
R5	3	0	3	0	0	3	0
R3 R5	3	3	0	3	0	0	3

Fig. 2 **(A)** A toy example of the constraint-based models. Rectangular nodes R1 to R7 are reactions, and the attached intervals represent lower and upper bounds of their reaction speeds. R1 is the nutrient uptake reaction. R6 is the cell growth reaction. R7 is the target metabolite production reaction. Circular nodes C1 to C4 are internal metabolites. **(B)** Reaction deletion strategies and the resulting flux (reaction speed) distributions.

to combinatorial explosion.

3. Results

We developed an algorithm DynCubeProd for calculating reaction deletion strategies that achieve growth coupling of designated target metabolite production in constraint-based models of metabolic networks.

Because DynCubeProd is a method obtained by improving CubeProd [9], in this section, we provide an overview of CubeProd and then explain the difference between DynCubeProd and CubeProd. The algorithm behavior of DynCubeProd is also illustrated using examples. The relationship between DynCubeProd and other methods is discussed in Section 4.

3.1 DynCubeProd

Idea: CubeProd considers the three-dimensional solution space whose axes represent GR, PR and sum of absolute values of fluxes (SF). Let TMGR, TMPR, and TMSF be the theoretical maximum values of the above, respectively. Then the whole constraint space is a rectangle formed by [0, TMGR], [0, TMPR], and [0,TMSF] [*1]. According to the designated value of P, each of [0, TMGR], [0, TMPR], and [0,TMSF] are divided into P pieces. Therefore, finally, CubeProd considers P^3 constraint sub-spaces.

The value of P closely affects the trade-off between the ease of finding a solution and the computation time. The larger the value of P, the easier it is to find a solution, but the slower the computation time. Therefore, if CubeProd uses a certain value of P and cannot find a solution, a larger value of P should be applied, but the computation time will be more. However, it may be the case that some of the small solution spaces generated by a larger P are already proved by a smaller P to contain no solutions.

DynCubeProd starts with P=1 and doubles P if no solution is found. When applying a larger P, it refers to the result of applying the smaller P and avoids searching for places where there is no solution.

Because the intervals on each of the three axes are equally subdivided into P sub-intervals, the entire constraint space is divided into P^3 sub-spaces, and

$$\frac{(i-1)\times TMGR}{P} \leq x_{growth} \leq \frac{i\times TMGR}{P},$$
$$\frac{(j-1)\times TMPR}{P} \leq x_{target} \leq \frac{j\times TMPR}{P},$$
$$\frac{(k-1)\times TMSF}{P} \leq \sum |x| \leq \frac{k\times TMSF}{P}$$

are added as constraints and the sum of the absolute values of fluxes is minimized for every $1 \leq i, j, k \leq P$, where i, j, k are integers.

In each of those P^3 sub-spaces, (1) LP is employed with the above three constraints, (2) if the LP is feasible, reactions whose flux is less than 10^{-5} are collected as K, (3)the minimum value of PR is calculated with deletions of K under the condition that GR is maximized without the above three constraints, and (4) if GR and PR exceed the minimum required values, the output K is considered as the solution.

The number of sub-spaces to be computed is P^3 and it increases dramatically as P increases, which will lead to a power-of-three increase in computation time. However, the larger the value of P, the smaller is the range of the subspace and the easier it is to approach the point of the optimal solution or local optimal solution.

A dynamic strategy is adopted by DynCubeProd to save time. Starting with $i = 1$ for $P = 2^i$, DynCubeProd increases i one by one, and stops once an acceptable knockout strategy is obtained. Suppose that Q is a sub-space corresponding to $P = m$ and $\{Q_1, \ldots, Q_k\}$ are sub-spaces of Q for $P = n$ with $m < n$. If the candidate knockout strategy computed from Q is not acceptable when $P = m$, all the sub-spaces $\{Q_1, \ldots, Q_k\}$ will be skipped during the calculation when $P = n$.

[*1] Because it is difficult to determine TMSF in polynomial time, we approximate TMSF by SF when PR is maximized, and [0,2·TMSF] was used instead of [0,TMSF] in the computational experiments and examples.

The example of results of DynCubeProd applied to the network of Fig. 2 (A) with $P = 1$ and $P = 2$ is published in https://doi.org/10.3389/fbinf.2021.716112 [4].

For a positive integer k, the solution obtained by DynCubeProd for $P = 2k$ is also a solution for $P = k$, and the constraint sub-spaces skipped by DynCubeProd do not include a solution.

Define S as the constraint space of an LP problem. Let $S_1 \cup S_2 \cup \cdots \cup S_n = S$ hold. Suppose there exists a solution x in the sub-space S_k, $x \in S_k$. Then, x must be in the space of S, $x \in S$, because $S_k \subset S$. If there is no solution x in S, that is, $x \notin S$, then such x must not exist in any sub-space of S.

3.2 Pseudo Code of DynCubeProd

The details of pseudo code is published in https://doi.org/10.3389/fbinf.2021.716112 [4].

3.3 Computational Experiments

The dataset used in the computational experiments was iJO1366, which is a GSMM of *Escherichia coli* K-12 MG1655 from the BiGG database with 1805 metabolites and 2583 reactions [3], [6]. All procedures of DynCubeProd were implemented based on Gurobi, COBRA Toolbox and Matlab on a Windows machine with Intel(R) Core(TM) i5-8500 CPU 3.00 GHz 6-core processor and 32.0 GB RAM.

If the target metabolite is not connected to an external reaction, then, an auxiliary external reaction is added, and the growth coupling is evaluated by GR and the outgoing flux from the additional external reaction, which is also called PR.

Fig. 3 (A) shows the computing time of DynCubeProd and CubeProd when applied to iJO1366 under aerobic conditions at different values of P. It also shows the ratio of the number of success cases to the number of target metabolites. For $P \geq 16$, the reaction deletion strategies were obtained for more than 95% of the target metabolites. It should be noted that the success ratio of DynCubeProd and CubeProd is always the same for the same P. The computing time for DynCubeProd with P=32 was only approximately quarter of the computing time of CubeProd with P=16. Fig. 3 (B) visually compares the computing time increase by P between DynCubeProd and CubeProd.

Furthermore, under anaerobic conditions of iJO1366, DynCubeProd succeeded in computing the reaction deletion strategies for 76 of the 211 target metabolites for

A

P	DynCubeProd time(s)	CubeProd time(s)	Success rate
2	557	630	74.93%
4	1012	4183	88.82%
8	2712	33262	93.76%
16	11888	264649	95.25%
32	62443		95.99%

B

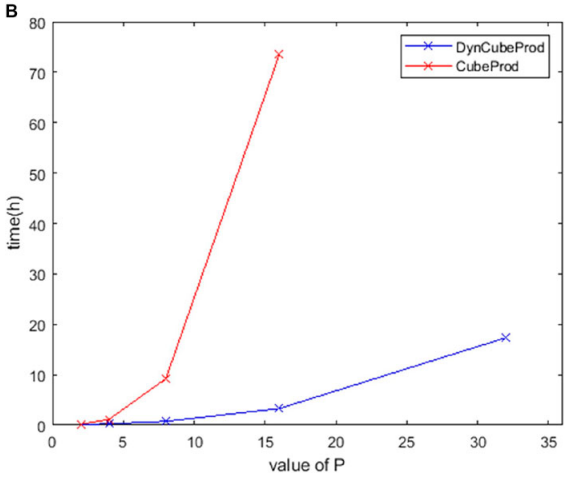


Fig. 3 **(A)** Computation time and success ratio when DynCubeProd and CubeProd were applied to iJO1366 under aerobic conditions for different values of P. **(B)** Visual comparison of the computation time of DynCubeProd and CubeProd of (A).

which the EFV-based method [11], which is one of the best methods, could not.

4. Discussion

The details of discussion part is published in https://doi.org/10.3389/fbinf.2021.716112 [4].

5. Conclusion

DynCubeProd is an improved version of CubeProd, which is an existing algorithm based on solution space decomposition. The improvements in DynCubeProd are as follows. (1) While CubeProd divides the solution space based on pre-specified parameters, DynCubeProd gradually divides the solution space into smaller and smaller pieces. (2) CubeProd mechanically explores even the subspace where no solution is expected to exist, while DynCubeProd stops dividing the solution space when no solution exists. (3) While CubeProd searches the entire solution space, DynCubeProd stops as soon as it finds a solution that meets the conditions. The results of computer experiments using iJO1366 confirmed that DynCubeProd reduces the computation time more than 10 times than CubeProd. The reduction in computation time enabled finer solution space partitioning, and reaction deletion strategies could be calculated for about 40% of the target metabolites for which reaction deletion strategies could

not be obtained by the EFV-based method. In this study, we developed DynCubeProd, by improving the computation speed of CubeProd, which enabled us to calculate reaction deletion strategies in anaerobic conditions for many target compounds that could not be calculated before.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

MY implemented the source codes and conducted computational experiments. TT designed the study. The manuscript was written both by MY and TT.

Funding

This study was partially supported by grants from JSPS, KAKENHI #20H04242.

Acknowledgments

We appreciate all members of the mathematical bioinformatics group of Bioinformatics Center, Institute for Chemical Research, Kyoto University.

参考文献

[1] Bailey, J. E.: Toward a science of metabolic engineering, *Science*, Vol. 252, No. 5013, pp. 1668–1675 (1991).
[2] Burgard, A. P., Pharkya, P. and Maranas, C. D.: Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization, *Biotechnology and bioengineering*, Vol. 84, No. 6, pp. 647–657 (2003).
[3] King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. and Lewis, N. E.: BiGG Models: A platform for integrating, standardizing and sharing genome-scale models, *Nucleic acids research*, Vol. 44, No. D1, pp. D515–D522 (2016).
[4] Ma, Y. and Tamura, T.: Dynamic solution space division-based methods for calculating reaction deletion strategies for constraint-based metabolic networks for substance production: DynCubeProd, *Frontiers in Bioinformatics*, p. 24.
[5] Maranas, C. D. and Zomorrodi, A. R.: *Optimization methods in metabolic networks*, John Wiley & Sons (2016).
[6] Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M. and Palsson, B. Ø.: A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011, *Molecular systems biology*, Vol. 7, No. 1, p. 535 (2011).
[7] Stephanopoulos, G., Aristidou, A. A. and Nielsen, J.: Metabolic engineering: principles and methodologies, *Metabolic Engineering Principles & Methodologies* (1998).
[8] Tamura, T.: Grid-based computational methods for the design of constraint-based parsimonious chemical reaction networks to simulate metabolite production: GridProd, *BMC bioinformatics*, Vol. 19, No. 1, p. 325 (2018).
[9] Tamura, T.: Efficient reaction deletion algorithms for redesign of constraint-based metabolic networks for metabolite production with weak coupling, *IPSJ Transactions on Bioinformatics*, Vol. 14, pp. 12–21 (2021).
[10] Tamura, T.: L1 Norm Minimal Mode-Based Methods for Listing Reaction Network Designs for Metabolite Production, *IEICE Transactions on Information and Systems*, Vol. 104, No. 5, pp. 679–687 (2021).
[11] von Kamp, A. and Klamt, S.: Growth-coupled overproduction is feasible for almost all metabolites in five major production organisms, *Nature communications*, Vol. 8, p. 15956 (2017).