Vol.2015-BIO-43 No.5
2015/9/12

Integer Programming-based Method for Designing Synthetic Metabolic Networks by Minimum Reaction Insertion in a Boolean Model

Wei Lu[1] Takeyuki Tamura[2,a)] Jiangning Song[3,4] Tatsuya Akutsu[2]

Abstract: In this technical report, we consider the Minimum Reaction Insertion (MRI) problem for finding the minimum number of additional reactions from a reference metabolic network to a host metabolic network so that a target compound becomes producible in the revised host metabolic network in a Boolean model. Although a similar problem for larger networks is solvable in a flux balance analysis (FBA)-based model, the solution of the FBA-based model tends to include more reactions than that of the Boolean model. However, solving MRI using the Boolean model is computationally more expensive than using the FBA-based model since the Boolean model needs more integer variables. Therefore, in this study, to solve MRI for larger networks in the Boolean model, we have developed an efficient Integer Programming formalization method in which the number of integer variables is reduced by the notion of feedback vertex set and minimal valid assignment. As a result of computer experiments conducted using the data of metabolic networks of *E. coli* and reference networks downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, we have found that the developed method can appropriately solve MRI in the Boolean model and is applicable to large scale-networks for which an exhaustive search does not work. We have also compared the developed method with the existing connectivity-based methods and FBA-based methods, and show the difference between the solutions of our method and the existing methods. Our developed software is available at "http://sunflower.kuicr.kyoto-u.ac.jp/~rogi/minRect/minRect.html".

1. Introduction

Metabolism is one of the most important biological processes in organisms. Relations between reactions and chemicals in the metabolism are often represented by metabolic networks. Since many of these metabolic processes can produce commodity and specialty chemicals, the manipulation of metabolisms has been extensively studied in the field of metabolic engineering. One of the most successful applications of metabolic engineering is production of industrially valuable products using a microbial host with recombinant technologies [1]. Techniques for production of desired chemicals using a microbial host are roughly classified into the following three types [2]: (a)combinations of existing pathways, (b)engineering of existing pathways, and (c)*de novo* pathway design. In (a), partial pathways can be recruited from independent organisms and co-localized in a single host. For example, 1,3-propanediol is synthesized by Nakamura *et al.* in which pathways from *Saccharomyces cerevisiae* and *Klebsiella pneumonia* were assembled *in E. coli* [3] and another example is the production of artemisinic acid, a precursor to the plant-based anti-malarial drug artemisinin in yeast [4]. In (b), new non-natural chemicals can be produced by engineering existing routes [5]. (c) is realized by the combination of (a) and (b), that is, the recruitment of partial pathways from different species and the use of engineered enzymes for extensions of pathways. It is to be noted that (a) focuses on the topology of the given metabolic networks, while (b) and (c) utilize the information of the structures of chemicals as well.

In the type (a) problem, it seems that there are three major models for judging the producibility of target compounds, that is, *connectivity model*, *flow model*, and *Boolean model*. For each of them, Minimum Reaction Insertion (MRI) problem can be defined for finding the minimum number of additional reactions from a reference metabolic network to a host metabolic network so that a target compound becomes producible in the revised host metabolic network. In the connectivity model such as [6], the producibility of target compounds is judged by the connectivity between the source and the target compounds. After the source and the target compounds are connected by the additional reactions, the producibility is often evaluated by such a flow model as flux balance analysis (FBA) or an elementary mode [7], in which the sum of incoming flows must be equal to the sum of outgoing flows for each compound and the ratio of the amount of substrates and products must satisfy the coefficients given in each chemical reaction formula. In the Boolean model, each reaction occurs if

[1] National Institute of Informatics, Tokyo 101-8430, Japan
[2] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
[3] Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia
[4] National Engineering Laboratory for Industrial Enzymes, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China
[a)] tamura@kuicr.kyoto-u.ac.jp

all its substrates are producible whereas each compound is producible if one of its producing reactions occurs [8]. The source compounds are called *seeds* and the producible compounds are called the *scope* of the seed. In this model, a Boolean function of "AND" is attached to each reaction node and "OR" is attached to each compound node in the metabolic networks.

For example, suppose that there is a chemical reaction "A + B → C + D", where A and B are called *substrates* whereas C and D are called *products*. In the connectivity model, either A or B is necessary to produce C and D, whereas both A and B are necessary for the Boolean model. In the flow model including FBA, in addition to the condition that both A and B must exist, both C and D are necessary to be consumed by other reactions. Thus, each model outputs a different solution for producing desired compounds.

From the view point of computational complexity, although the connectivity model is very simple and then applicable even to very large networks, its logical analysis ability is not strong since it cannot detect the lack of necessary substrates. The good point of the flow model is its computational efficiency since problems in the flow model can often be formalized by linear programming, for which there exist polynomial time algorithms [9]. However, these polynomial time algorithms are not applicable for MRI since discrete variables are necessary for representing additional reactions, although it is solvable by mixed integer programming [10].

Although the computational time of the FBA-based method for MRI is very small and scalable for genome-scale metabolic reconstruction [10], Boolean methods also have attractive features and are expected to complement the FBA-based method. Indeed, for the analysis of metabolic networks, many studies have been conducted to develop Boolean models. For example, Lemke *et al.* [11] studied the effect of deletion of each enzyme in the metabolic network of a Boolean model, and Smart *et al.* [12] considered almost the same problem from the viewpoint of the Boolean aspect of the flux balance model. Li *et al.* [13] and Sridhar *et al.* [14] have developed methods for finding a set of enzymes whose inhibition stops the production of the target compounds with a minimum elimination of the non-target compounds. Lee *et al.* [15] and Takemoto *et al.* [16] estimated the distribution of the size of the effect of the deletions of enzymes using a branching process.

As for the shortcoming of the FBA-based method for MRI, it tends to be considerably affected by the redundancy of the given metabolic network since each node is affected not only by the incoming flows but also by the outgoing flows. For example, suppose that a metabolic network of Fig. 1 (A) is given, where circles and rectangles represent compounds and reactions respectively. In order to produce the target compound from the source compounds, {R1, R2, R3, R4} is necessary in the flow model including FBA, whereas either {R1, R4} or {R1, R2, R3} is sufficient for the Boolean model. Moreover, in the metabolic network of Fig. 1 (B), {R1,R2,R3} is necessary for FBA whereas {R2} is sufficient for the Boolean model.

Therefore, in this research, we study the problem of designing a pathway for producing target compounds in metabolic networks of the Boolean model since its logical analysis ability is more stable than that of the FBA, particularly when the flexible parts of the metabolic networks are large. Our approach is based on (a), that is, the combination of existing pathways. In our problem setting, a base metabolic network of a host organism, which we call the *host network*, is given; it cannot produce the target compound in its initial form. However, an integrated metabolic network of many other organisms are given as the *reference network* from which we should find the minimum number of additional reactions so that the target compound becomes producible. We prove that this problem is NP-complete.

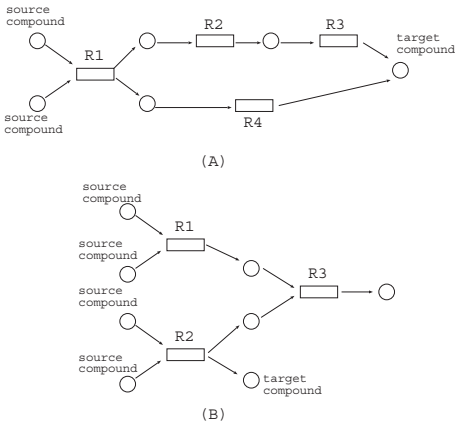

Fig. 1 A problem of how to produce a target compound from the source nodes. In the Boolean model, either {R1, R4} or {R1, R2, R3} is sufficient, whereas {R1, R2, R3, R4} is necessary for the flow model including FBA.

Although both the FBA-based model and the Boolean model for MRI are considered to be NP-complete, the former is likely to have a faster exponential time algorithm than the latter since FBA has fewer integer variables. Although the computational complexity of the Boolean model is large, we develop an efficient method based on integer programming (IP) [17], [18], which is often used as a formalization of NP-complete problems and there is an efficient free solver for IP called CPLEX [19]. We also conducted four computer experiments in which the metabolic network of *E. coli* is used as the host network and the reference pathway of the KEGG database [20] is used as the reference network, and propanol, butanol, sedoheptulose 7-phosphate, and maleic acid are used as the target compound in each experiment. The results of the experiments show that (1) our IP-based method can appropriately solve MRI in the Boolean model; (2) solutions of MRI in the Boolean model are more suitable than those by connectivity based methods; (3) our IP-based method is applicable to large-scale networks where an exhaustive search does not work; and (4) solutions of MRI in the Boolean model tend to be smaller than those in the FBA-based model based on [21]. Our developed software is available at "http://sunflower.kuicr.kyoto-u.ac.jp/˜rogi/minRect/minRect .html"

Since the full version is available as [22], this technical report partially omits the details.

2. Materials and Methods

2.1 Problem Definition

In this section, the main problem **Minimum Reaction Insertion (MRI)** in a Boolean model is first explained with an example

and then mathematical formalization is described.

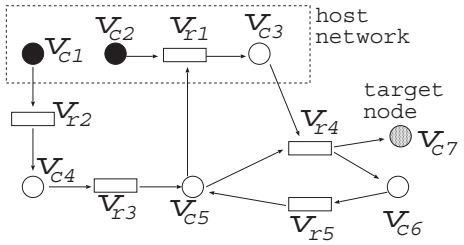

Fig. 2 An example of MRI. v_{c_1} and v_{c_2} are the source nodes.

Suppose that a metabolic network shown in Fig. 2 is given, where each rectangle (resp., circle) corresponds to a reaction (resp., chemical compound). For example, v_{r4} is a reaction, its substrates are v_{c3} and v_{c5} and its products are v_{c6} and v_{c7}. Black circles v_{c1} and v_{c2} denote the source nodes and are assumed to be provided by the external environment. On the other hand, a gray circle v_{c7} represents a target compound and the purpose of MRI is to make the target compound producible. However, initially only the host network, which is shown by the dotted rectangle, is available. Since only v_{c_1}, v_{c_2}, v_{c3} and v_{r1} are included in the host network, the target compound v_{c7} is not producible. Instead the entire network is called the reference network and reactions not included in the host network can be added later. In MRI, the minimum number of additional reactions should be determined to make the target compound producible. In this example, the addition of $\{v_{r2}, v_{r3}, v_{r4}\}$ is the optimal solution. The difficult point of MRI is how to deal with the effect of cycles. In the example of Fig. 2, the addition of $\{v_{r4}, v_{r5}\}$ looks like the optimal solution. However, this solution is not appropriate since it relies on the cycle consisting of $\{v_{c6}, v_{r5}, v_{c5}, v_{r4}\}$ and v_{c7} is not producible unless the initial amount of v_{c6} is sufficiently large.

MRI is mathematically defined as follows: A *metabolic network* can be represented by a directed graph $G = (V, E)$. There are two types of node sets V_c and V_r, where V_c denotes a set of *compound nodes* and V_r represents a set of *reaction nodes*. $V = V_c \cup V_r$ and $V_c \cap V_r = \{\}$ hold. The neighbors of compound nodes must be reaction nodes, and the neighbors of reaction nodes must be compound nodes. Let $V_s \subseteq V_c$ be a set of *source nodes* and $v_t \in V_c$ be a *target node*. Source nodes have no incoming edges and correspond to the seed compounds of [8]. In this study, we assume that source nodes are producible at any time.

Suppose that a *host network* $G_1 = (V_1, E_1)$ and a *reference network* $G_2 = (V_2, E_2)$ are given where G_1 and G_2 are metabolic networks, and G_1 is a subgraph of G_2 induced by V_1. V'_c (resp., V'_r) is a set of compound nodes (resp., reaction nodes) in $V_2 - V_1$ and is called the set of *additional compound nodes* (resp., *additional reaction nodes*).

Let $V_a \subseteq V'_r$ be a set of additional reaction nodes. In the Boolean model, each node is assigned either "0" or "1". For a compound node, "1" means producible and "0" means not producible. As for a reaction node, "1" means active and "0" means inactive. Let A be such an assignment (that is A is a function from V to $\{0, 1\}$). For each node $v \in V$, we write $v = 0$ (resp., $v = 1$) if 0 (resp., 1) is assigned to v. A is called a *valid assignment* if the following conditions are satisfied: (i) for each $v \in V_s$, $v = 1$. (ii) for each $v \in V_c - V_s$, $v = 1$ if and only if there is u such that $(u, v) \in E$ and $u = 1$. (iii) for each $v \in V_r$, $v = 1$ if and only if $v \in V_a \cup V_1$ and $u = 1$ holds for all u such that $(u, v) \in E$. This implies that each reaction node corresponds to an "AND" node and each compound node corresponds to an "OR" node.

If G_2 has no directed cycles, a valid assignment is uniquely determined for each V_a. However, if G_2 has a directed cycle, multiple valid assignments may exist. Let us call $v_i \in V_s$ and $v_j \in V_c - V_s$ *source connected* if there is a directed path from v_i to v_j, and the values of the nodes included in the path are all 1. There exist valid assignments where the values of nodes in a directed cycle are 1 even if these nodes are not source connected. In order to avoid such a case, we use the notion of *minimal valid assignment*, which is similar to the notion of maximal valid assignment defined in [23]. A valid assignment A is called *minimal* if A is valid and $\{v | v = 1, v \in V\}$ is minimal with respect to the inclusion relationships for sets.

Now we define the **Minimum Reaction Insertion** as follows:

- **Input:** A host metabolic network $G_1 = (V_1, E_1)$, a reference metabolic network $G_2 = (V_2, E_2)$, and a target compound v_t.
- **Output:** A minimum cardinality set of V_a for which $v_t = 1$ is satisfied in the minimal valid assignment of the induced subgraph of G_2 by $V_1 \cup V'_c \cup V_a$.

Although it is not described in this version, a minimal valid assignment is uniquely determined if V_a is given. However, solving MRI is not easy since the number of candidate V_a is $2^{|V'_r|}$ and MRI is proved to be NP-complete. Since utilizing software packages of Integer Programming (IP) is efficient for solving NP-complete problems, we develop a method of IP formalization for solving MRI. Since the computational time of the IP-based method is considered to be exponential in terms of the number of variables, it is important to develop an IP formalization of MRI with a small number of variables. To do so, our previously developed method for **Minimum Reaction Cut** (MRC) [23] may be useful although many modifications are necessary.

MRC is a problem to find a minimum set of reactions that interfere with the production of target compounds [23] and is known to be NP-complete. Let m (resp., n) be the number of compound (resp., reaction) nodes. If we use $m + n$ time steps to calculate the maximal valid assignment in MRC, the number of variables in IP is $O((m+n)^2)$. The feedback vertex set (FVS) is a node set whose removal makes a network cycle-free. In [23], we succeeded in reducing the number of variables to $O(f(f+m+n))$, where f is the size of the feedback vertex set and f is considerably smaller than m or n. If use of $O((m+n)^2)$ variables is allowed in MRI, almost the same method as in MRC can be used. However, to reduce the number of variables in IP to $O(f(f+m+n))$, many modifications are necessary since minimal valid assignment and maximal valid assignment have different features.

2.2 Integer Programming-Based Method for Minimum Reaction Insertion

Here, we show IP formalization methods for MRI in the Boolean model. To apply IP, problems must be formalized to maximize or minimize a given objective function which is a linear

function of integer variables and constraints must also be given as linear equations or inequations of integer variables.

Suppose that the host network and the reference network are given as shown in Fig. 2. The simplest IP formalization **IP-MRI-A** for solving **Minimum Reaction Insertion** is as follows where the time step increases by 1 when the Boolean calculation is synchronously conducted for every node:

IP-MRI-A

Minimize

TER2(0)+ TER3(0)+ TER4(0)+ TER5(0) (1)

Subject to

TC7(m+n) = 1 (2)

for all $t = 0, \ldots, m+n$

TR1(t+1) + FC2(t) + FC5(t) ≥ 1,
FR1(t+1) + TC2(t) ≥ 1,
FR1(t+1) + TC5(t) ≥ 1 (3)

TR2(t+1) + FC1(t) + FER2(t) ≥ 1,
FR2(t+1) + TC1(t) ≥ 1,
FR2(t+1) + TER2(t) ≥ 1 (4)

TR3(t+1) + FC4(t) + FER3(t) ≥ 1,
FR3(t+1) + TC4(t) ≥ 1,
FR3(t+1) + TER3(t) ≥ 1 (5)

TR4(t+1) + FC3(t) + FC5(t) + FER4(t) ≥ 1,
FR4(t+1) + TC3(t) ≥ 1,
FR4(t+1) + TC5(t) ≥ 1,
FR4(t+1) + TER4(t) ≥ 1 (6)

TR5(t+1) + FC6(t) + FER5(t) ≥ 1,
FR5(t+1) + TC6(t) ≥ 1,
FR5(t+1) + TER5(t) ≥ 1 (7)

TC3(t+1) = TR1(t) (8)

TC4(t+1) = TR2(t) (9)

FC5(t+1) + TR3(t) + TR5(t) ≥ 1,
TC5(t+1) + FR3(t) ≥ 1,
TC5(t+1) + FR5(t) ≥ 1 (10)

TC6(t+1) = TR4(t) (11)

TC7(t+1) = TR4(t) (12)

TER2(t+1) = TER2(t),
TER3(t+1) = TER3(t),
TER4(t+1) = TER4(t),
TER5(t+1) = TER5(t) (13)

TC1(t) = 1, TC2(t) = 1 (14)

TC3(0) = TC4(0) = TC5(0) =
TC6(0) = TC7(0) = 0 (15)

TR1(0) = TR2(0) = TR3(0) =
TR4(0) = TR5(0) = 0 (16)

TX + FX = 1 for any X (17)

where every variable takes either 0 or 1. $v_{ri} = 1$ (resp., $v_{ri} = 0$) at time step t is represented by TR$i(t)$=1 (resp. FR$i(t)$=1) and TR$i(t)$+FR$i(t)$=1 holds for any i and t. For example, TR2(1)=0 means that $v_{r2} = 0$ at time step 1, and FR2(1)=1 automatically holds at the same time. In the implementation, FR$i(t)$ is replaced with 1-TR$i(t)$ to reduce the number of variables. Similarly, the values of compound nodes are represented by TC$i(t)$ and FC$i(t)$. For example, FC4(3) = 1 means that $v_{c4} = 0$ at time step 3.

(3) represents the Boolean relation $v_{r1}(t+1) = v_{c2}(t) \wedge v_{c5}(t)$. Since Boolean relations such as "$\wedge$" or "$\vee$" cannot directly be used in IP, it is necessary to convert them into linear equations and/or inequations. Since $x_1 = x_2 \wedge x_3 \wedge \cdots \wedge x_k$ can be represented by $(x_1 \vee \overline{x_2} \vee \overline{x_3} \vee \cdots \vee \overline{x_k}) \wedge (\overline{x_1} \vee x_2) \wedge (\overline{x_1} \vee x_3) \wedge \cdots \wedge (\overline{x_1} \vee x_k) = 1$, $v_{r1}(t+1) = v_{c2}(t) \wedge v_{c5}(t)$ can be converted into $(v_{r1}(t+1) \vee \overline{v_{c2}(t)} \vee \overline{v_{c5}(t)}) \wedge (\overline{v_{r1}(t+1)} \vee v_{c2}(t)) \wedge (\overline{v_{r1}(t+1)} \vee v_{c5}(t)) = 1$, and then (3) is obtained.

For a compound node with indegree 1, the value of the predecessor node is just copied. For example, since v_{c3} has only one predecessor v_{r1}, $v_{c3}(t+1)$ is just copied from $v_{r1}(t)$ as shown in (8). Similarly, $v_{c4}(t+1)$ is just copied from $v_{r2}(t)$ as shown in (9).

For a compound node with indegree more than 1, it is necessary to convert the "$\vee$" relation into linear equations or equations. (10) represents the Boolean relation $v_{c5}(t+1) = v_{r3}(t) \vee v_{r5}(t)$. Since $x_1 = x_2 \vee x_3 \vee \cdots \vee x_k$ is represented by $(\overline{x_1} \vee x_2 \vee x_3 \vee \cdots \vee x_k) \wedge (x_1 \vee \overline{x_2}) \wedge (x_1 \vee \overline{x_3}) \wedge \cdots \wedge (x_1 \vee \overline{x_k}) = 1$, $v_{c5}(t+1) = v_{r3}(t) \vee v_{r5}(t)$ can be converted into $(v_{c5}(t+1) \vee \overline{v_{r3}(t)} \vee \overline{v_{r5}(t)}) \wedge (\overline{v_{c5}(t+1)} \vee v_{r3}(t)) \wedge (\overline{v_{c5}(t+1)} \vee v_{r5}(t)) = 1$, and then (10) is obtained.

As for the reaction nodes not included in the host network, TER$i(t)$ and FER$i(t)$ are used to represent whether v_{ri} is activated. We use a virtual node v_{ei} as one of the predecessors of v_{ri}. Since v_{ri} is represented by an AND node, $v_{ei} = 0$ keeps v_{ri} inactive even if all other predecessors of v_{ri} are 1. For example, v_{r2} in Fig. 2 has only one predecessor v_{c1}. However, since v_{r2} is not included in the host network and $v_{ei} = 1$ is necessary for $v_{r2} = 1$, $v_{r2}(t+1) = v_{c1}(t) \wedge v_{e2}(t)$ must hold, and then (4) is obtained.

Since we assume minimal valid assignment, at $t = 0$, the source compound nodes are assigned 1, but the other compound nodes and reaction nodes are assigned 0.

$m+n$ is the largest number of time steps necessary for the 0-1 assignment to converge. (1) means that the number of additional reactions should be minimized. (2) means that the target compound v_{c7} should become 1 after the 0-1 assignment converges. (3)-(7) represent the constraints by v_{r1} to v_{r5} respectively. Note that v_{e1} does not exist since v_{r1} is included in the host network and then $v_{r1} = 1$ holds for any V_a. (8)-(12) represent the constraints by v_{c3} to v_{c7} respectively. (13) represents that V_a does not change by time transition. (14) means that v_{c1} and v_{c2} are source nodes. (15)-(16) represent that all nodes but source nodes are assigned 0 in the initial state. (17) means that "T" and "F" represent "true (1)" and "false (0)" respectively, and complement each other.

The above formalization can clearly solve MRI and obtain the correct solution $V_a = \{v_{r2}, v_{r3}, v_{r4}\}$, however $O((m+n)^2)$ variables are necessary. To reduce the number of variables, it is necessary to reduce the number of time steps. If time is not taken into account at all, an inappropriate IP formalization **IP-MRI-B** is obtained. Details of IP-MRI-B is described in [22].

In IP-MRI-B, the solution of IP is $V_a = \{v_{r4}, v_{r5}\}$ since $(v_{r1}, \ldots, v_{r5}, v_{c1}, \ldots, v_{c7}) = (1,0,0,1,1,1,1,1,0,1,1,1)$ is a valid assignment and satisfies $v_{c7} = 1$. Note that v_{r2} and v_{r3} are forced to be 0 since they are not included in either the host network or V_a. Although it satisfies all constraints and $|V_a|$ is minimum, this assignment is not appropriate since $\{v_{r4}, v_{c6}, v_{r5}, v_{c5}\}$ forms a cycle and all of them are assigned 1 without the influence of source nodes. To avoid such an inappropriate assignment, it is necessary to consider minimal valid assignment with respect to the number

of 1s for each V_a. Although it is not described in this version, the minimal valid assignment is uniquely determined for each V_a.

Thus, IP-MRI-A can solve MRI, but $m + n$ time steps are necessary, while IP-MRI-B, which does not use the notion of time, cannot solve MRI. The feedback vertex set (FVS) is a set of nodes whose removal makes the network acyclic. Since IP-MRI-B can solve MRI if there is no cycle, it is reasonable to apply IP-MRI-B for the acyclic network obtained by the deletion of FVS and use the notion of time as in IP-MRI-A to nodes included in F based on the idea developed in [23].

In the improved method, IP-MRI-C, we firstly find an FVS F consisting of reaction nodes and then decompose each $v_{ri} \in F$ into two nodes v_{ri} and v_{si} so that v_{ri} has only in-edges and v_{si} has only out-edges. For example, in the network of Fig. 2, since $F = \{v_{r4}\}$ is a feedback vertex set, v_{r4} is decomposed into v_{r4} and v_{s4} as shown in Fig. 3. Furthermore, we put an additional constraint that $v_{si}(t + 1) = v_{ri}(t)$. The number of time steps of IP-MRI-C is $f + 1$ while that of IP-MRI-A is $m + n + 1$, where $f = |F|$. Therefore, the numbers of variables in IP-MRI-C and IP-MRI-A are $O(f(m + n + f))$ and $O((m + n)^2)$ respectively. Since f is considerably smaller than $m + n$ in most metabolic networks and the computational time of IP exponentially increases with the number of variables, we can expect a significant improvement from the view point of the computational time.

Although finding the minimum FVS is known to be NP-complete, it is not necessary to use the minimum FVS in our problem setting. To choose FVS, we use a simple greedy algorithm GreedyFVS, that is described in [22].

Since the reaction nodes for FVS are chosen by a greedy algorithm, the size of FVS is not always optimal. However, it is important to note that even if the size of FVS is not optimal, the solution of MRI calculated by IP-MRI-C is always optimal. If there are multiple optimal solutions in MRI, there is a possibility that the solutions are different since IP outputs only one solution. However, it may be possible to enumerate all optimal solutions of MRI by iteratively solving IP with a constraint to avoid the already chosen solutions.

The example of IP-MRI-C for Fig. 2 is explained in [22].

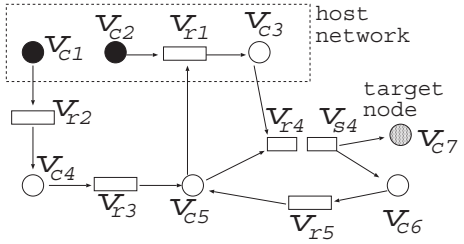

Fig. 3 The cycles are decomposed in the FVS-based method.

3. Results

3.1 Computer Experiments

We conducted computer experiments for solving MRI with data downloaded from the KEGG database. The experiment was conducted on a PC with an Intel(R) Xeon(R) 3.33GHz CPU and 10GB RAM having the SUSE Linux (version 12.2) operating system, where CPLEX (version 12.4.0.0) was used as the solver of integer programming.

In this study, a reference network consists of the central metabolism and the related modules necessary for producing the target compound. A map of the KEGG PATHWAY is a minimum unit, and three or four maps of the KEGG PATHWAY are chosen and integrated as the reference network in each of our experiments. As for species, a reference network includes the chemical reactions of all species, whereas the metabolic networks of *E. coli* are used for the host networks. The major difference between the pathway alignment methods by KEGG and our developed method is that our method is based on a Boolean model, whereas the pathway alignment methods consider only the topology of networks.

In synthetic biology, it is of great interest to construct a minimal genome that realizes the desired functions [24], [25], [26]. Since glycolysis, gluconeogenesis, citrate cycle and pentose phosphate pathway are considered to be essential even in artificial organisms, it is reasonable to assume that the host networks in the computer experiments have some of these pathways in one of the simplest organisms, *E. coli*. Because the purpose of this study is not focused on the reconstruction of genome-scale metabolic network model, but the design of a minimal genome in addition to the existing pathways to produce a desired compound, each reference network consists of the maps of the KEGG pathway located between the central metabolism and each target compound.

In the first computer experiment, the target compound is propanol (C00479 in KEGG ID), the host network is glycolysis and gluconeogenesis of *E. coli* (eco00010.xml), and the reference network covers glycolysis, gluconeogenesis and glycerolipid metabolism of other species (ko00010.xml and ko00561.xml). The numbers of compound and reaction nodes are 58 and 85, respectively, where 30 reactions are reversible. The source nodes are D-glucose (C00031), oxaloacetate (C00036), salicin (C01451), arbutin (C06186), UDP-glucose (C00029), acyl-CoA (C00040), and diglucosyl-diacylglycerol (C06040). It took 0.19 s to solve MRI. The obtained additional reactions are V_a = {R01514, R01752, R01036, R01048, R02577, R02376}, where these reactions produce propanol from 3-phospho-D-glycerate (C00197) via glycerol (C00116). Since 3-phospho-D-glycerate (C00197) is producible by glycolysis and gluconeogenesis of *E. coli* and works as a connection between glycolysis and glycerolipid metabolism, the obtained V_a can be considered an appropriate solution of MRI.

3.2 Difference between Developed Model and Shortest Path-Based Model

To show the difference between the developed model and the shortest path-based models, we conducted the second experiment where PathComp of KEGG (”http://www.genome.jp/tools/pathcomp/”) was used to calculate the solution of the shortest path-based model. In the experiment, the host network consists of glycolysis, gluconeogenesis and citrate cycle of *E. coli* (eco00010.xml and eco00020.xml), and the reference network consists of glycolysis, gluconeogenesis, citrate cycle and pentose phosphate pathway of other species (ko00010.xml, ko00020.xml and ko00030.xml). The numbers of compound and reaction nodes are 64 and

108, respectively, where 59 reactions are reversible. There are four source nodes, D-glucose(C00031), arbutin(C06186), salicin(C01451), and acetate (C00033), and the number of candidates for the additional reactions is 66. When the target compound is sedoheptulose 7-phosphate (C05382), the solution of MRI is V_a = {R01827, R01830}, where the substrates of R01827 are beta-D-fructose 6-phosphate (C05345) and D-erythrose 4-phosphate (C00279). It took 32.58 s to obtain the solution. Since D-erythrose 4-phosphate (C00279) is not included in the host network, it is necessary to add R01830 in which substrates are beta-D-fructose 6-phosphate (C05345) and D-glyceraldehyde 3-phosphate (C00118) and the products are D-xylulose 5-phosphate (C00231) and D-erythrose 4-phosphate (C00279). It is to be noted that both beta-D-fructose 6-phosphate (C05345) and D-glyceraldehyde 3-phosphate (C00118) are producible by the host network.

On the other hand, PathComp just connects the producible compounds and the target compound adds only R01827 since R01827 is adjacent to both beta-D-fructose 6-phosphate (C05345) and sedoheptulose 7-phosphate (C05382). However, it is clear that R01827 does not occur if D-erythrose 4-phosphate (C00279) does not exist. Thus the difference between the shortest path-based method and the developed method is that the developed method considers Boolean constraints for each reaction and compound whereas the shortest path-based method only considers the connectivity of nodes.

3.3 Scalability

Next, we conducted the third experiment to show the scalability of our method. The host network consists of the source nodes of glycolysis and gluconeogenesis of *E. coli* (eco00010.xml), that is, D-glucose(C00031), arbutin(C06186), salicin(C01451), oxaloacetate(C00036) and acetate (C00033). The reference network consists of glycolysis, gluconeogenesis, citrate cycle, pentose phosphate pathway and butanol metabolism of other species (ko00010.xml, ko00020.xml, ko00030.xml and ko00650.xml), where R01172 is treated as a reversible reaction. The target compound is butanol (C06142). The numbers of compound and reaction nodes are 93 and 150, respectively, where 87 reactions are reversible. It took 919.79 s (15m19s) for the developed method to solve MRI and the solution was V_a = {R00235, R00238, R01977, R03027, R01171, R01172, R03545}. These seven reactions form a path from acetate to 1-butanol via acetyl-CoA, acetoacetyl-CoA, crotonoyl-CoA and butanoyl-CoA, which satisfies the Boolean constraints. Since the number of reactions in the reference network is 150, it is necessary to examine ${}_{150}C_7$ cases if an exhaustive search is conducted. Since examining ${}_{150}C_7 \simeq 2.941 \times 10^{11}$ cases is almost impossible, it is seen that the IP-based method is useful for solving MRI, particularly when the given networks are not small.

3.4 Difference between Developed Model and FBA-Based Model

Finally, we conducted an experiment to show the difference between the developed model and the FBA-based model. We assume that the reference network consists of glycolysis, gluconeogenesis, citrate cycle, pentose phosphate pathway and butanol metabolism of other species (ko00010.xml, ko00020.xml, ko00030.xml and ko00650.xml), and the host network includes only one reaction R04394 between salicin (C01451) and salicin 6-phosphate (C06188). Therefore, the source node is only salicin (C01451). Note that reversible reactions are decomposed into two reactions, and denoted by P and Q. The target compound is maleic acid (C01384). The numbers of compound and reaction nodes are 93 and 150, respectively, where 87 reactions are reversible.

Then, the solution of MRI in our Boolean model is {R05134, R02736, R02035, R02036, R05605, R00344, R00342, R01082, R01087}, whereas the solution of FBA-based model is {R05134, R02736, R02035, R02036, R05605, R01058, R01518, R00658, R00200, R00344, R00342, R01082, R01087}. It is to be noted that {R01058, R01518, R00658, R00200} is not necessary for the Boolean model, but necessary for the FBA-based model. In the Boolean model, R01058 is not necessary to produce C01384 since the lack of reactions in downstream does not affect. However, in the FBA model, R01058 is necessary. Otherwise, C00118 is not consumed and then R05605 cannot occur. Thus, the solution of MRI in the FBA-based model tends to include more reactions than that in the Boolean model. It took 7896.46 s (2h11m36s) to solve the Boolean model of MRI.

4. Discussion

In this technical report, we formalized an optimization problem MRI in a Boolean model with a notion of minimal valid assignment. We proved that MRI in the Boolean model is NP-complete and the minimal valid assignment is uniquely determined when V_a is given. Since an exhaustive search cannot be used to solve MRI when the given networks are not small, we developed an IP-based method for MRI. To improve the scalability of the developed method, it is necessary to reduce the number of variables appearing in IP formalization since the computational time of IP is considered to be exponential to the number of variables. Although the simple IP formalization with the notion of time is useful for solving MRI, it needs $O((m+n)^2)$ variables in IP formalization. If the notion of FVS is used, the number of necessary time steps reduces to f, where f denotes the size of FVS, and the number of variables in IP is $O(f(m+n+f))$. Although the idea of using FVS is similar to [23], many modifications are necessary since the minimal valid assignment and the maximal valid assignment have many different properties.

We also conducted four computer experiments in which data were downloaded from the KEGG database, CPLEX was used as the IP solver, and propanol, butanol, sedoheptulose 7-phosphate, and maleic acid were used as the target compound for each experiment. The host network was a metabolic network of *E. coli* and the reference network of KEGG was used as the reference network. The results of the computer experiments confirmed the correctness and the scalability of the developed method, and the appropriateness of the problem setting of MRI.

An important advantage of our Boolean model is its capability of detecting the lack of substrates, whereas the connectivity-based methods cannot appropriately handle this point. An extended type of connectivity-based method is BNICE, which enu-

merates all possible pathways from the source nodes to the target compound, and uses thermodynamical feasibility and pathway length to evaluate each candidate pathway. In contrast, the developed method evaluates each candidate pathway based on the number of additional reactions. Another advantage of the developed model is its capability of handling branches and/or cycles in a pathway from the source compounds to the target compound, whereas BNICE considers only the non-branching paths. However, since BNICE nicely evaluates each pathway by the thermodynamic free energy of the included compounds and length, considering the thermodynamic free energy in a Boolean model represents an important direction of our future work.

It is to be noted that the solution of MRI in the FBA-based model is different from that in the Boolean model. In particular, if the reference network includes a large redundant part, the FBA-based model tends to output a larger solution than the Boolean model, although the FBA-based model is very fast when compared to the Boolean model. Therefore, one of our future works is to develop a hybrid method combining the FBA-based method and the Boolean-based method.

References

[1] Bro C, Regenberg B, Forster J, Nielsen J (2006) In silico aided metabolic engineering of saccharomyces cerevisiae for improved bioethanol production. Metabolic Engineering 8(2): 102-111.

[2] Prather K, Martin C (2008) De novo biosynthetic pathways: rational design of microbial chemical factories. Current Opinion in Biotechnology 19(5): 468–474.

[3] Nakamura CE, Whited GM (2003) Metabolic engineering for the microbial production of 1,3-propanediol. Current Opinion in Biotechnology 14(5): 454 - 459.

[4] Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, et al. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature 440(7086): 940-943.

[5] de Boer AL, Schmidt-Dannert C (2003) Recent efforts in engineering microbial cells to produce new chemical compounds. Current Opinion in Chemical Biology 7(2): 273 - 278.

[6] Noor E, Eden E, Milo R, Alon U (2010) Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. Molecular Cell 39(5): 809 - 820.

[7] Haus UU, Klamt S, Stephen T (2008) Computing knock-out strategies in metabolic networks. Journal of Computational Biology 15(3): 259-268.

[8] Handorf T, Ebenhoh O, Heinrich R (2005) Expanding metabolic networks: Scopes of compounds, robustness, and evolution. Journal of Molecular Evolution 61(4): 498-512.

[9] Karmarkar N (1984) A new polynomial-time algorithm for linear programming. In: Proceedings of the sixteenth annual ACM symposium on Theory of computing. New York, NY, USA: ACM, STOC '84, pp. 302–311. DOI: 10.1145/800057.808695. URL http://doi.acm.org/10.1145/800057.808695.

[10] Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology 9: 977-982.

[11] Lemke N, Herédia F, Barcellos CK, Dos Reis AN, Mombach JC (2004) Essentiality and damage in metabolic networks. Bioinformatics 20(1): 115–119.

[12] Smart AG, Amaral LAN, Ottino JM (2008) Cascading failure and robustness in metabolic networks. Proceedings of the National Academy of Sciences 105(36): 13223-13228.

[13] Li Z, Wang RS, Zhang XS, Chen L (2009) Detecting drug targets with minimum side effects in metabolic networks. Systems Biology, IET 3(6): 523-533.

[14] Sridhar P, Song B, Kahveci T, Ranka S (2008) Mining metabolic networks for optimal drug targets. In: Pacific Symposium on Biocomputing. volume 13, pp. 291–302.

[15] Lee D, Goh KI, Kahng B (2012) Branching process approach for boolean bipartite networks of metabolic reactions. Physical Review E 86(2): 027101.

[16] Takemoto K, Tamura T, Akutsu T (2013) Theoretical estimation of metabolic network robustness against multiple reaction knockouts using branching process approximation. Physica A: Statistical Mechanics and its Applications 392(21): 5525–5535.

[17] Schrijver A (1986) Theory of linear and integer programming. New York, NY, USA: John Wiley & Sons, Inc.

[18] Li Z, Zhang S, Wang Y, Zhang XS, Chen L (2007) Alignment of molecular networks by integer quadratic programming. Bioinformatics 23(13): 1631-1639.

[19] (2010). IBM ILOG CPLEX Optimizer. urlhttp://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

[20] Kanehisa M, Goto S (2000) Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1): 27-30.

[21] Suthers PF, Dasika MS, Kumar VS, Denisov G, Glass JI, et al. (2009) A genome-scale metabolic reconstruction of *mycoplasma genitalium*, *i*ps189. PLoS Computational Biology 5(2).

[22] Lu W, Tamura T, Song J, Akutsu T (2014) Integer programming-based method for designing synthetic metabolic networks by minimum reaction insertion in a Boolean model, PLoS ONE 9(3).

[23] Tamura T, Takemoto K, Akutsu T (2010) Finding minimum reaction cuts of metabolic networks under a boolean model using integer programming and feedback vertex sets. IJKDB 1(1): 14-31.

[24] Lee JH, Sung BH, Kim MS, Blattner FR, Yoon BH, et al. (2009) Metabolic engineering of a reduced-genome strain of escherichia coli for l-threonine production. Microb Cell Fact 8(2).

[25] Ara K, Ozaki K, Nakamura K, Yamane K, Sekiguchi J, et al. (2007) Bacillus minimum genome factory: effective utilization of microbial genome information. Biotechnology and Applied Biochemistry 46(3): 169–178.

[26] Mizoguchi H, Mori H, Fujio T (2007) Escherichia coli minimum genome factory. Biotechnology and Applied Biochemistry 46(3): 157–167.