2008－BIO－14　(4)
2008／9／18

整数計画と帰還点集合による代謝ネットワークの構造的堅牢性の測定

田村武幸 [1]　竹本和広 [2]　阿久津達也 [1]
[1] 京都大学　[2] 東京大学

概要　代謝ネットワークの堅牢性は、標的化合物を少なくとも一つ合成不能にする為に阻害しなければならない酵素の最少数で測定することができる。我々はこの問題に対し、帰還点集合 (FVS) を用いて閉路を処理することにより従来の整数計画に基づく方法を改良した。大腸菌の解糖/糖新生、クエン酸回路、ペントースリン酸経路に提案手法を適用したところ、阻害すべき酵素の最適集合を見つけるのに要した時間は 0.23～5.15 秒であり、従来の方法に比べ 7.63～145.82 倍高速であった。

Measuring Structural Robustness of Metabolic Networks Using Integer Programming and Feedback Vertex Sets

Takeyuki Tamura[1], Kazuhiro Takemoto[2], and Tatsuya Akutsu[1]
[1] *Kyoto University,*　[2] *University of Tokyo*

Abstract Robustness in metabolic networks can be measured by the minimum number of enzymes to be disrupted so that at least one of the target compounds cannot be synthesized. We developed an improved integer programming-based method for this problem using feedback vertex sets (FVS) to cope with cycles. When applied to E. coli metabolic network consisting of Glycolysis/Gluconeogenesis, Citrate cycle and Pentose phosphate pathway, the proposed method could find an optimal set of enzymes to be disrupted in 0.23~5.15 sec, which are 7.63~145.82 times faster than the previous method.

1 Introduction

In order to understand the principles of living organisms, it is quite important to reveal the origin of the robustness. Burgard et al. developed *bilevel programming* methods to find strategies for maximizing or minimizing the production of target chemical compound by knockout of a limited number of genes or reactions [1, 2]. In their methods, the problems are first formalized as bilevel integer programs and then these are transformed into conventional integer programs by making use of the duality in linear programming. Their methods are elegant and scalable, and are very close to the purpose of this article. In order to effectively apply flux balance analysis, we should know exact or near exact network topology and stoichiometry parameters. However, details of such information may not be available in some cases. Furthermore, flux balance based methods assume that their exist steady-states, which may not exist in some cases. Thus, it is worthy to develop another approach. Ruths et al. recently proposed a combinatorial approach for analyzing signaling networks [3]. They proved NP-hardness of the problem and showed some heuristic algorithm. In their model, it is allowed that both chemical compounds and reactions are inactivated. However, it seems difficult to inactivate specific chemical compounds, and thus their model does not seem to be appropriate for analyzing the robustness of metabolic networks.

In order to study structural and combinatorial aspects of metabolic networks, Handorf et al. recently introduced the concept of *scope* [4]. The scope is a set of all possible metabolites obtained from a given set of *seed* compounds and a given structure of a metabolic network. Handorf et al. also defined the inverse of the scope problem: given a set of metabolites, compute or enumerate minimal sets of seed compounds [5]. Furthermore, they studied robustness of scopes against modifications of enzymes [4]. However, the methods for analyzing robustness are based on exhaustive or random disruptions of enzymes and can not be directly used for measuring the structural robustness of a metabolic network.

Based on the concept of scope, we previously studied computational aspects of robustness of metabolic networks [6]. We defined this problem as follows: given a metabolic network, a set of seed compounds and a set of target compounds, determine the minimum number of reactions whose deletion prevents generation of some of the target compounds (i.e., the set of target compounds is no more a subset of the scope). We proved that this problem is NP-hard. We developed an $O(1.822^n)$ time algorithm for a special case in which the number of substrates per reaction is at most two, where n is the number of reactions. We also developed an *integer programming* (IP) based method for a general case. However, the number of variables appearing in an integer program was $O((m+n)^2)$ where m is the number of compounds and thus the method could not be applied to large-scale networks.

In this article, we significantly improve the IP-based method by means of a novel use of a *feedback vertex set* (FVS). A FVS is a concept in graph theory and is a set of vertices removal of which makes a given graph cycle-free. It is well-known that computation of the minimum cardinality FVS is NP-hard for both directed and undirected graphs, whereas several approximation algorithms have been developed [7, 8]. In our purpose, it is not necessary to use the optimal FVS. Thus, we adopted a simple greedy algorithm to compute an FVS. By using an FVS, we can reduce the number of variables in IP from $O((m+n)^2)$ to $O(f(f+m+n))$ where f is the size of an FVS. Since f is usually very small (e.g., $10 \sim 20$ in our experiments) and IP's computational time usually grows exponentially in the number of variables, it leads to a significant improvement of practical efficiency. When applied to E. coli metabolic network consisting of Glycolysis/Gluconeogenesis, Citrate cycle and Pentose phosphate pathway extracted from KEGG database, the improved IP method was 7.63~145.82 times faster than our previous IP method.

2 Methods

2.1 Problem Definition

Here we review the definition of the problem [6]. Let $V_c = \{v_{c_1}, \ldots, v_{c_m}\}$ and $V_r = \{v_{r_1}, \ldots, v_{r_n}\}$ be a set of *compound nodes* and a set of *reaction nodes* respectively, where $V_c \cap V_r = \{\}$. It is to be noted that most reactions are catalyzed by enzymes and thus each of most reactions can be inactivated by disruption of a gene corresponding to the enzyme catalyzing the reaction. Let $V = V_c \cup V_r$. Let $V_s \subseteq V_c$ and $V_t \subseteq V_c$ are a set of *source nodes* and a set of *target nodes* respectively, where $V_s \cap V_t = \{\}$.

A *metabolic network* is defined as a directed graph $G(V, E)$ satisfying the following conditions: (i) For each edge $(u, v) \in E$, either $(u \in V_c) \wedge (v \in V_r)$ or $(u \in V_r) \wedge (v \in V_c)$ holds. (ii) Each source node does not have an incoming edge. (iii) Each target node does not have an outgoing edge. (iv) Each node $v \notin V_s$ has at least one incoming edge.

Let $V_a \subseteq V_r$ be a set of reaction nodes corresponding to a set of inactivated reactions. We assign 0-1 value to each node V. Let A be such an assignment (i.e., 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 $v \notin V_a$ holds and $u = 1$ holds for all u such that $(u, v) \in E$.

The second condition means that compound nodes correspond to OR nodes. The third condition means that reaction nodes correspond to AND nodes, where the output is forced to be 0 if a node is inactive (i.e., a node belongs to V_a). Clearly, the following proposition holds.

Proposition 1 *[6] If there is no directed cycle in $G(V, E)$, a valid assignment is uniquely determined for a given V_a.*

If there exist directed cycles, we may have multiple valid assignments. However, we can define the maximal valid assignment uniquely. A valid assignment A is called *maximal* if A is valid and $\{v | v = 1, v \in V\}$ is maximal.

Proposition 2 *[6] A maximal valid assignment is determined uniquely for a given metabolic network $G(V, E)$ and a given set of inactivated reactions V_a.*

For example in Fig. 1 (A), assume that $V_c = \{v_{c_1}, \ldots, v_{c_8}\}$, $V_r = \{v_{r_1}, v_{r_2}, v_{r_3}\}$, $V_s = \{v_{c_1}, v_{c_3}, v_{c_6}\}$ and $V_t = \{v_{c_8}\}$ are given. The maximal valid assignment for $V_a = \{v_{r_2}\}$ is $\{v_{c_1}, \ldots, v_{c_8}, v_{r_1}, v_{r_2}, v_{r_3}\}=\{1,1,1,1,1,1,0,1,1,0,1\}$. $\{v_{c_1}, \ldots, v_{c_8}, v_{r_1}, v_{r_2}, v_{r_3}\}=\{1,0,1,0,0,1,0,0,0,0,0\}$ is not maximal but valid for $V_a = \{v_{r_2}\}$. Intuitively, maximal valid assignment is the steady state to which a metabolic network converges assuming all nodes were initially assigned 1 and nodes in V_a were inactivated.

Let K_{in} be the maximum indegree among reaction nodes (i.e., $|\{u | (u, v) \in E\}| \leq K_{in}$ holds for all $v \in V_r$). Let K_{out} be the maximum outdegree among reaction nodes (i.e., $|\{u | (v, u) \in E\}| \leq K_{out}$ holds for all $v \in V_r$). It is reasonable to assume that K_{in} and K_{out} are bounded by a constant since the number of substrate compounds and the number of product compounds are bounded by a constant.

Then, we formulate the problem of determining the robustness of a metabolic network (**MetaboRobust**) as follows:

Input: A metabolic network $G(V, E)$.

Output: A minimum cardinality set V_a for which $v = 0$ is satisfied for some $v \in V_t$ in the maximal valid assignment.

It is to be noted that we only consider irreversible reactions for the simplicity. Reversible reactions can be taken into account if we replace one reversible reaction with two irreversible reactions. In such a case, special treatment is needed since inactivation of single reversible reactions corresponds to inactivation of two irreversible reactions.

In MetaboRobust, a set of target compounds is given. But, it is almost equivalent to the case where only one target compound is given because the original problem can be reduced to $|V_t|$ cases of the single target problem. In highly robust cells such as cancer cells, it may not be enough to prevent generation of a single compound in V_t because there may exist hidden or unknown reactions. In such a case, it may be useful to prevent generation of all compounds in V_t. We also consider this variant, which is referred as **MetaboRobustAll** and **MetaboRobustII-All**. Note that target nodes may have outgoing edges for MetaboRobustAll.

2.2 Improved IP Formalization

In the formalization in [6], the number of variables of IP is $O((m+n)^2)$. So, this method cannot be applied to large scale networks. Furthermore, real networks contain many reversible reactions. In such a case, it may output inadequate solutions as follows. Assume that a reversible reaction of Fig. 1 (B) is given. If the maximal valid assignment is defined as the solution, both v_{c_1} and v_{c_2} are assigned 1 and they never become 0 unless v_{r_1} is inactivated. If v_{c_1} or v_{c_2} is provided by another reaction, this constraint is reasonable. However, if neither v_{c_1} nor v_{c_2} is provided by another reaction, it is more reasonable to assume that v_{c_1} and v_{c_2} eventually attenuate and become $v_{c_1} = v_{c_2} = 0$. Therefore we have to extend the definition of MetaboRobust as follows. Let $v_{s_1}, \ldots, v_{s_{k_1}}$ and $v_{p_1}, \ldots, v_{p_{k_2}}$ are substrates and products of a reversible reaction where products

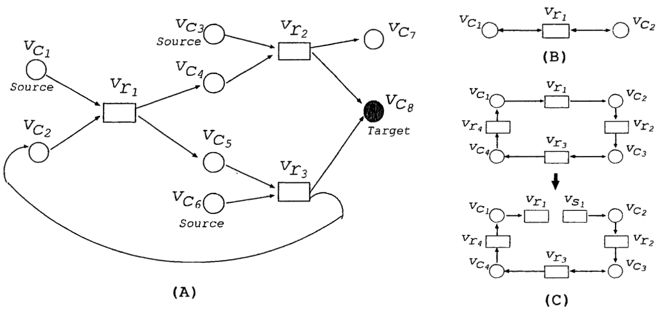

Figure 1: (A) $V_a = \{v_{r_1}\}$ is the solution of MetaboRobust when $V_c = \{v_{c_1}, \ldots, v_{c_8}\}$, $V_r = \{v_{r_1}, v_{r_2}, v_{r_3}\}$, $V_s = \{v_{c_1}, v_{c_3}, v_{c_6}\}$ and $V_t = \{v_{c_8}\}$ are given. (B)Example of reversible reaction. (C)Example of decomposition of cycle.

can exchange for substrates and vice versa. An assignment A is called extended maximal valid assignment if A is valid, $\{v | v = 1, v \in V\}$ is maximal, and both v_{s_i} and v_{p_j} are assigned 1 only if either all v_{s_i} or all v_{p_j} are provided by the other reactions. Now, we formulate the extended version of MetaboRobust.

Problem: MetaboRobustII:

Input: A metabolic network $G(V, E)$.

Output: A minimum cardinality set V_a for which $v = 0$ is satisfied for some $v \in V_t$ in the extended maximal valid assignment.

It is not straightforward to solve this problem by IP. However, by utilizing the notion of FVS F, the problem can be solved in most cases. Recall that an FVS is a set of nodes removal of which makes the network acyclic. Recall also that we need not introduce the notion of time if the network is acyclic. Thus, we combine these two ideas. In the improved method, we first identify an FVS that consists of reaction nodes, and then apply the IP formulation similar to that of MetaboRobust to the obtained acyclic network. Different from the previous IP-formulation for cyclic cases, we use the same time step for all the nodes in the obtained acyclic network. In order to take the effect of cycles into account, we create another reaction node v_{s_i} for each reaction node v_{r_i} in F (see Fig. 1 (C)) and put an additional constraint that $v_{s_i}(t+1) = v_{r_i}(t)$. By means of this modification, the number of required time steps is reduced from $m + n + 1$ to $f + 1$, where $f = |F|$. Therefore, the required number of variables in IP-formulation is reduced from $O((m+n)^2)$ to $O(f(m+n+f))$. Since it is expected that f is much smaller than $m + n$ and the computation time of IP exponentially increases with the number of variables, it is a significant improvement. Though finding an FVS of minimum cardinality is NP-hard, we need not use an FVS of minimum cardinality. Thus, we employ a simple greedy method to select an FVS. If there does not exist a reversible reaction, this FVS-based method should output the same robustness measure as the previous IP formulation for cyclic cases does.

3 Results and Discussion

In this article, we deal with two versions of the main problem, **MetaboRobust** and **MetaboRobustII**, and their definitions are different from each other. The difference between them are mainly caused by how to treat reversible reactions. In MetaboRobust, every compound is assumed to be producible at an initial state and deletions of some reactions gradually affect the whole network and eventually a target compound becomes nonproducible. However, in this definition, compounds directly connected to a reversible reaction v_{r_1} never become nonproducible unless v_{r_1} is deleted. For example, in Fig. 1 (B), v_{c_1} and v_{c_2} are directly connected to a reversible reaction v_{r_1}. Since v_{c_1} and v_{c_2} are assigned 1 at an initial state, they never become nonproducible unless v_{r_1} is deleted. However, it is more reasonable to assume that v_{c_1} and v_{c_2} become nonproducible if neither v_{c_1} nor v_{c_2} is provided by another reaction. Thus in MetaboRobustII, the extended version of MetaboRobust, every compound is assumed to be producible at an initial state, but it will become nonproducible if it is not producible only from seed compounds.

We implemented Integer Programming based methods for these two problems, where their formalizations are explained in Method Section, for E. coli metabolic network consisting of Glycolysis/Gluconeogenesis (00010), Citrate cycle (00020) and Pentose phosphate pathway (00030) from KEGG database. It contains 60 compound nodes and 111 reaction nodes (44 reversible reactions and 23 irreversible reactions). Reversible reactions are represented by two nodes so that each edge has only one direction. Note that $44 \times 2 + 23 = 111$. Pyruvate (C00022), Acetyl-CoA (C00024), Acetate (C00033), Oxaloacetate (C00036) and Phosphoenolpyruvate (C00074) were used as target compounds from a view point of importance of amino acids. The experiment was done on a PC with Xeon 3GHz CPUs and 8GB RAM under the Linux (version 2.6.24) operating system, where CPLEX (Version 10.1.0) was used as the solver of integer programming.

Elapsed times for MetaboRobust and MetaboRobustII for each target compound are shown in Table 1. When the target compound is Pyruvate (C00022), Acetyl-CoA (C00024), Acetate (C00033), Oxaloacetate (C00036) , Phosphoenolpyruvate (C00074) respectively, the elapsed time of the computational experiment for MetaboRobust was 10.15, 46.88, 49.93, 42.62, 65.62 seconds respectively, whereas those for MetaboRobustII

Table 1: Elapsed time for MetaboRobust and MetaboRobustII for each target compound.

Target compound	Computational time for MetaboRobust	Computational time for MetaboRobustII	Ratio
C00022	10.15s	0.23s	44.13
C00024	46.88s	4.39s	10.68
C00033	49.93s	4.95s	10.09
C00036	42.62s	4.91s	8.68
C00074	65.62s	0.45s	145.82
MetaboRobustAll	39.28s	5.15s	7.63
Number of variables in IP	81396	6526	12.22

Table 2: Solution for MetaboRobustII for each target compound

Target compound	Indegree	The number of deleted reactions	Deleted reactions
C00022	2	2	R00200, R05605
C00024	4	2	R00351, R07618
C00033	2	2	R00351, R007618
C00036	4	3	R00351, R01518, R02570
C00074	2	2	R00341, R01518
MetaboRobustAll		4	R00351,R01518,R02570,R05605

was 0.23, 4.39, 4.95, 4.91, 0.45 respectively. Since ratios of them are 44.13, 10.68, 10.09, 8.68 and 145.82, it is seen that solving MetaboRobustII is much faster than solving MetaboRobust. Furthermore, for the problem where all of Pyruvate (C00022), Acetyl-CoA (C00024), Acetate (C00033), Oxaloacetate (C00036) and Phosphoenolpyruvate (C00074) should become nonproducible, which we call **MetaboRobustAll**, elapsed times for MetaboRobust and MetaboRobustII are 39.28 and 5.15 respectively. Since the ratio is 7.63, solving MetaboRobustII was much faster than solving MetaboRobust.

The reason why MetaboRobustII is faster than MetaboRobust is that numbers of variables in IP in MetaboRobustII is much less than that of MetaboRobust. In MetaboRobust, we have to use 81396 variables whereas only 6526 variables were used in MetaboRobustII and their ratio is 7.63. To deal with cycles included in a given network, we have to introduce the notion of time. In MetaboRobust, the number of used time steps is $m + n + 1$. Since each node requires $O(m + n)$ variables, $O((m + n)^2)$ variables are necessary for IP formalization.

On the other hand, in MetaboRobustII, only f time steps are necessary for each node, where f is the cardinality of FVS of a given network. However, each node v_{r_i} included in the FVS is decomposed into two nodes v_{r_i} and v_{s_i} so that v_{r_i} has only in-edges and v_{s_i} has only out-edges. Therefore, since the number of nodes becomes $f + m + n$, the total number of variables in IP is $O(f(f+m+n))$, where f was 13 in our computational experiment. This is the reason why solving MetaboRobustII is much faster than solving MetaboRobust.

The minimum cardinality set of deleted reactions for MetaboRobust and MetaboRobustII for each target compound is shown in Table 2.

4 Conclusions

In this article, we introduced a novel Integer Programming formalization method utilizing FVS for finding the minimum cardinality set of reactions whose deletion causes that a target compound becomes nonproducible. We applied the proposed method to E. coli metabolic network consisting of Glycolysis/Gluconeogenesis, Citrate cycle and Pentose phosphate pathway from KEGG database. Pyruvate, Acetyl-CoA, Acetate, Oxaloacetate and Phosphoenolpyruvate were used as target compounds from a view point of importance of amino acids. The result of the computational experiments showed that our proposed method can appropriately find the solution of MetaboRobustII since there were good agreement with the existing knowledge of metabolic networks. Furthermore, the elapsed time for the computational experiment was much faster than the method used in [6].

References

[1] Burgard, A. P., Pharkya, P., Maranas, D.: OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnology and Bioengineering, 84. 647–657 (2003).

[2] Dashika, M. S., Burgard, A., Maranas, D.: A computational framework for the topological analysis and targeted disruption of signal transduction networks. Biophysical Journal, 91, 382–398 (2006).

[3] Ruths, D. A., Nakhleh, L., Iyengar, M. S., Reddy, S. A. G., Ram, P. T.: Hypothesis generation in signaling networks, Journal of Computational Biology, 13, 1546–1557 (2006).

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

[5] Handorf, T., Christian, N., Ebenhöh, O., Kahn, D.: An environmental perspective on metabolism, Journal of Theoretical Biology, in press.

[6] Tamura, T., Takemoto, K., Akutsu, T.: Algorithms for measuring structural robustness of metabolic networks. Submitted.

[7] Bar-Yehuda, R., Geiger, D., Naor, J., Roth, R. M.: Approximation algorithms for the feedback vertex set problem with applications to constraint satisfaction and Bayesian inference. SIAM Journal on Computing, 27, 942–959 (1998).

[8] Even, G., Naor, J., Schieber, B., Sudan, M.: Approximating minimum feedback sets and multicuts in directed graphs. Algorithmica, 20, 151–174 (1998).