Predicting Gene Knockout Effects by Minimal Pathway Enumeration

Takehide Soh
Transdisciplinary Research Integration Center
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan
soh@nii.ac.jp

Katsumi Inoue
National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, Japan
ki@nii.ac.jp

Tomoya Baba
Transdisciplinary Research Integration Center
1111 Yata, Mishima, 411-8540, Japan
tobaba@lab.nig.ac.jp

Toyoyuki Takada, Toshihiko Shiroishi
National Institute of Genetics
1111 Yata, Mishima, 411-8540, Japan
{ttakada,tshirois}@lab.nig.ac.jp

***Abstract*—In this paper, we propose a method to predict gene knockout effects for the cell growth by utilizing biological databases such as KEGG and EcoCyc, in which biological knowledge and experimental results have been collected. We construct biological networks from such databases and configure experimental conditions by giving source metabolites, target metabolites, and knockout genes. We then enumerate all minimal active pathways, which are minimal subsets of a given network using source metabolites to produce target metabolites. We simulate the effects of gene knockouts by measuring the difference of minimal active pathways between original networks and knockout ones. In the experiments, we applied it to predict the gene knockout effects on the glycolysis pathway of *Escherichia coli*. In the results, our method predicted three out of four essential genes, which are confirmed by the Keio collection containing comprehensive cell growth data obtained from biological experiments.**

Keywords-metabolic pathways; gene knockout; prediction method; minimal pathway; Keio collection.

I. Introduction

Living organisms, such as bacteria, fishes, animals, and humans, are kept alive by a huge number of intracellular chemical reactions. In *systems biology*, interactions of such chemical reactions are represented in a network called a *pathway*. Pathways have been actively researched in the last decade [1]–[3]. In addition, it is a biologically important subject to reveal the function of genes, which affect the phenotype of organisms. For model organisms such as *Escherichia coli* (*E. coli*), it has been approached by various methods. Constructing gene knockout organisms is an example of such methods [4]–[6]. However, it generally involves high costs and is limited by target genes and organisms.

In this paper, we propose a computation method to predict gene knockout effects by identifying *active pathways*, which are sub-pathways that produce target metabolites from source metabolites. We particularly focus on *minimal active pathways*, which are proposed by Soh and Inoue [7] and do not contain any other active pathways. In other words, all elements of each minimal active pathway are qualitatively essential to produce target metabolites. To predict gene knockout effects by the enumeration of minimal active pathways, we first introduce *extended pathways* that include relations between enzymatic reactions and genes. Then, we formalize the problem of finding minimal active pathways on the extended pathway with gene knockouts. After computing the solution of the problem, our method predicts gene knockout effects by collecting minimal active pathways that are still active under given gene knockouts.

To evaluate our method, we choose *E. coli* as our target organism, since it has been studied and much information about it is available on public resources. We apply our method to predict gene knockout effects on *E. coli* utilizing biological databases KEGG and EcoCyc, in which biological knowledge and experimental results have been collected. In the experiments, we compared our prediction and the cell growth of every single gene knockout *E. coli* strain, which was obtained from the Keio collection [4].

This paper is organized as follows. At first, we explain databases used in this paper and our research framework in Section II. We define the extended pathway in Section III. We formalize the problem of finding minimal active pathways on the extended pathway in Section IV and the effect of gene knockouts in Section V. Following that, we show our computational method in Section VI. In Section VII, we compare computational prediction and results of biological experiments. Following discussions in Section VIII, we conclude this paper in Section IX.

II. Used Databases and Research Framework

This section explains used databases and our research framework shown in Figure 1. In this paper, we particularly focus on *E. coli*. The metabolic pathway has been revealed by biochemical, molecular, and genetic studies, and *E. coli* is the organism in most detail. A large number of *E. coli* studies has contributed to several kinds of biological databases. In particular, we use the following two databases to construct our input network, called an extended pathway. One is *EcoCyc* [8]. It is a bioinformatics database that describes the genome and the biochemical machinery of *E. coli* K-12 MG1655. The EcoCyc project performs literature-based curation of the entire genome, metabolic pathways, etc. Specifically, it has been doing a literature-based curation

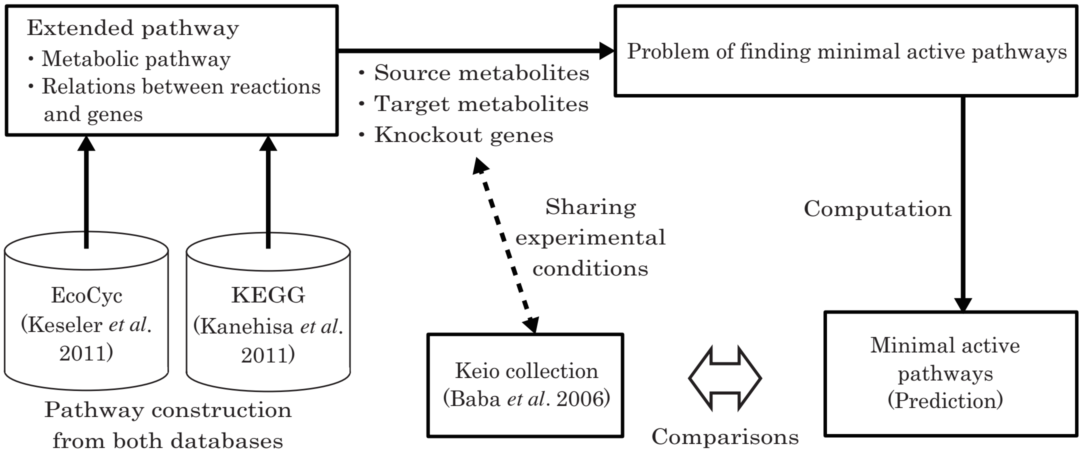
Extended pathway
• Metabolic pathway
• Relations between reactions and genes
• Source metabolites
• Target metabolites
• Knockout genes
Problem of finding minimal active pathways
Sharing experimental conditions
Computation
EcoCyc (Keseler et al. 2011)
KEGG (Kanehisa et al. 2011)
Pathway construction from both databases
Keio collection (Baba et al. 2006)
Comparisons
Minimal active pathways (Prediction)

Figure 1. Used Databases and our Research Framework

from more than 19,000 publications. We construct metabolic pathways with EcoCyc. The other one is *Kyoto encyclopedia of genes and genomes* (KEGG) [9], which is a database resource that integrates genomic, chemical, and systemic functional information. In particular, gene catalogs in the completely sequenced genomes, from bacteria to humans, are linked to higher-level systemic functions of the cell, the organism, and the ecosystem. A distinguished feature of KEGG is that it provides useful application program interfaces (API). We connect enzymatic reactions of metabolic pathways to genes with this API.

Figure 1 shows our research framework using the two databases. At first, we construct our input network called *extended pathway* from them. We then construct the problem of finding minimal active pathways by giving source and target metabolites to the extended pathway. In addition, the condition of knockout genes is also added to the problem. Then, we compute minimal active pathways using source metabolites to produce target metabolites. In the case of wild cells, we usually obtain multiple minimal active pathways including bypass pathways. However, in the case of knockout cells, we lose some (or all) of them. In brief, we predict the effects of gene knockouts from how many pathways are lost from the case of wild cells.

To evaluate our prediction method, we usually need additional biological experiments. However, Baba *et al.* comprehensively experimented on the cell growth of every single gene knockout strain [4]. Thanks to this research, we can evaluate our method with comparative ease. We briefly explain this research as follows. The *E. coli* K-12 single gene knockout mutant set, named *Keio collection*, is constructed as a resource for systems biological analyses. Excluding repetitive genes, e.g., insertion sequences related genes, 4288 protein coding genes are targeted for the systematic single gene knockout experiments. Of those, 3985 genes are successfully disrupted, and those of single-gene knockout mutants are constructed as the Keio collection. On the other hand, 303 genes are not disrupted and they are thought to be essential gene candidates. Those single gene knockout mutants have the same genome background, which results in an advantage for distinct functional analysis of the targeted gene. The genome-wide relationship between the genome structure, i.e., genotype, and the phenomena, i.e., phenotype, which are analyzed by using the Keio collection has become available.

Although Figure 1 shows specific databases for *E. coli*, the research framework itself can be applied for other organisms whose pathway information is available, e.g., mice.

III. Extended Pathways

In this section, we explain how to represent metabolic pathways and their relations to genes. We then define the extended pathway.

To represent metabolic pathways, we commonly use bipartite directed graph representation as follows. Let M be a set of metabolites and R be a set of reactions. For M and R, $M \cap R = \emptyset$ holds. Let $A_M \subseteq (R \times M) \cup (M \times R)$ be a set of arcs. A *metabolic pathway* is represented in a directed bipartite graph $\mathcal{G}_\mathcal{M} = (M \cup R, A_M)$, where M and R are two sets of nodes, and A_M is a set of arcs. In addition to the metabolic pathway, we consider relations between enzymatic reactions and genes. Let G be a set of genes and A_G be a set of arcs such that $A_G \subseteq (G \times R)$. That is, A_G represents relations between enzymatic reactions and genes. Let N be a set of nodes such that $N = M \cup R \cup G$ and A be a set of arcs such that $A = A_M \cup A_G$. Then, the *extended pathway* is represented in a directed graph $\mathcal{G} = (N, A)$.

Figure 2 shows an example of the extended pathway. As the figure shows, it consists of two layers: the metabolic layer and the genetic layer. The genetic layer is the difference between the metabolic pathway and the extended pathway. In this example, the pathway consists of nodes of $M = \{m_1, m_2, \ldots, m_6\}$, $R = \{r_1, r_2, \ldots, r_7\}$, and $G = \{g_1, g_2, \ldots, g_8\}$. Each arc represents relations between

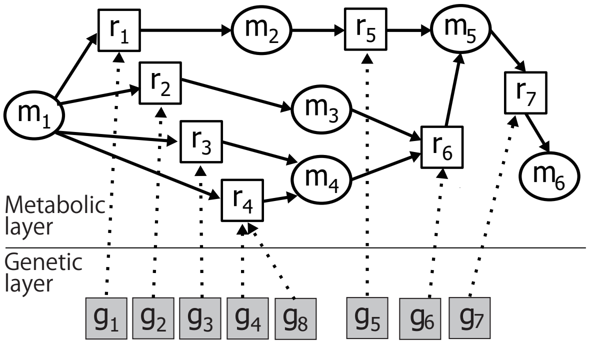

Figure 2. Example of an Extended Pathway

elements. For instance, the activation of the reaction r_6 needs the production of metabolites m_3 and m_4 and the expression of either g_4 or g_8. We will explain the interpretation of the extended pathway in detail in the next section.

IV. Minimal Active Pathways with Gene Knockouts

In the literature [7], the minimal active pathway is defined only on the metabolic pathway. On the other hand, in this section, we define the *minimal active pathway* on the extended pathway.

We here define $M_S \subset M$ as a set of source metabolites and $M_T \subset M$ as a set of target metabolites such that $M_S \cap M_T = \emptyset$. An *extended pathway instance* is represented in a four tuple $\pi = (N, A, M_S, M_T)$, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. Let K be a set of genes such that $K \subseteq G$. We use K as a set of knockout genes in a given pathway. A *knockout instance* is represented in a five tuple $\pi_K = (N, A, M_S, M_T, K)$. If $K = \emptyset$ then π_K corresponds to π.

Let m, r be a metabolite and a reaction such that $m \in M$ and $r \in R$, respectively. A metabolite $m \in M$ is called a *reactant* of a reaction $r \in R$ when there is an arc $(m, r) \in A$. On the other hand, a metabolite $m \in M$ is called a *product* of a reaction $r \in R$ when there is an arc $(r, m) \in A$. Furthermore, a gene $g \in G$ is called a *corresponding gene* of a reaction $r \in R$ when there is an arc $(g, r) \in A$.

A reaction is called a *reversible reaction* if it can occur in both directions between reactants and products. In this paper, we distinguish a reversible reaction as two reactions. Suppose that there is a reversible reaction r_1 that has m_1 and m_2 as reactants and m_3 and m_4 as products. In this case, we split the reaction r_1 into two reactions r_{1a} and r_{1b} such that one of them has m_1 and m_2 as products and m_3 and m_4 as reactants.

Let $s : R \rightarrow 2^M$ be a mapping from a set of reactions to a power set of metabolites such that $s(r) = \{m \in M \mid (m, r) \in A\}$ represents the set of metabolites that are needed to turn the reaction r activatable. Let $p : R \rightarrow 2^M$ be a mapping from a set of reactions to a power set of metabolites such that $p(r) = \{m \in M \mid (r, m) \in A\}$ represents the set of metabolites that are produced by the reaction r. Let $c : R \rightarrow 2^G$ be a mapping from a set of reactions to a power set of genes such that $c(r) = \{g \in G \mid (g, r) \in A\}$ represents the set of genes that are corresponding genes of the reaction r. Let $p' : M \rightarrow 2^R$ be a mapping from a set of metabolites to a power set of reactions such that $p'(m) = \{r \in R \mid (r, m) \in A\}$. Let $c' : G \rightarrow 2^R$ be a mapping from a set of genes to a power set of reactions such that $c'(g) = \{r \in R \mid (g, r) \in A\}$.

Let t be an integer variable representing time. In this paper, the time is used to represent order relation between reactions to produce target metabolites from source metabolites. In the following, we explain important notions related to production of metabolites, activation of reactions, and expression of genes. Since we focus on gene knockouts, we suppose that almost all genes exist in the cell of a given organism. We also suppose that if genes exist, then they are expressed and available to construct enzymes needed for enzymatic reactions. The reason for this condition is that we want to simulate how the lack of corresponding genes affects metabolic pathway rather than how the existence of genes affects other elements. Although our pathway modeling is simple, it allows us to analyze a whole cell scale pathway. Let $\pi_K = (N, A, M_S, M_T, K)$ be a knockout instance, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. Let $\mathcal{G} = (N, A)$ be an extended pathway. Let $M' \subset M$ be a subset of metabolites. A metabolite $m \in M$ is obviously *producible* at time $t = 0$ from M' on $\mathcal{G}$ if $m \in M'$ holds. A reaction $r \in R$ is *activatable* at time $t > 0$ from M' on $\mathcal{G}$ if the following two conditions are satisfied: (i) for every $m \in s(r)$, m is producible at time $t - 1$ from M', (ii) at least one corresponding gene $g \in c(r)$ is not included in K. A metabolite $m \in M$ is *producible* at time $t > 0$ from M' on $\mathcal{G}$ if there is at least one activatable reaction r at time t such that $m \in p(r)$. If r is activatable at time t, then r is activatable at time $t + 1$. If m is producible at time t, then m is producible at time $t + 1$.

Let $\mathcal{G}' = (N', A')$ be a sub-graph of $\mathcal{G}$, where $N' = M' \cup R' \cup G'$ and $A' = A'_M \cup A'_G$. Then, an active pathway of $\pi_K = (N, A, M_S, M_T, K)$ is defined as follows.

Definition 1: Active Pathway of Knockout Instance
A bipartite directed graph $\mathcal{G}'$ is an *active pathway* of π_K if it satisfies the following conditions:

- $M_T \subset M'$
- $M' = M_S \cup \{m \in M \mid (m, r) \subseteq A, r \in R'\} \cup \{m \in M \mid (r, m) \subseteq A, r \in R'\}$
- $A' = \{(m, r) \in A \mid r \in R'\} \cup \{(r, m) \in A \mid r \in R'\} \cup \{(g, r) \in A \mid g \notin K, r \in R'\}$
- $G' = \{g \in G \mid (g, r) \in A', r \in R'\}$
- For every $m \in M'$, m is producible from M_S on $\mathcal{G}'$

From Definition 1, active pathways include a set of metabolites, reactions, and genes, which are producible and activatable from M_S on $\mathcal{G}'$ such that all target metabolites M_T become producible. The number of active pathways depends on the combination of M_S and M_T but an extended pathway generally has a large number of active pathways. We thus particularly focus on minimal ones rather than active pathways. We give the definition of minimal active pathways of π_K as follows. Let $\mathcal{G}$ and $\mathcal{G}'$ be extended pathways. We say that $\mathcal{G}$ is *smaller* than $\mathcal{G}'$ and represented in $\mathcal{G} \subset \mathcal{G}'$ if $R \subset R'$. An active pathway $\mathcal{G}$ is *minimal active pathway* of π_K *iff* there is no active pathway of π_K, which is smaller than $\mathcal{G}$. As this definition shows, we only need to see sets of reactions to compare two pathways. Thus, in the rest of this paper, we sometimes represent a minimal active pathway as a set of reactions.

Any reactions included in a minimal active pathway cannot be deleted to produce target metabolites. Intuitively, this indicates that each of the elements of a minimal active pathway is essential. In practice, minimal active pathways including a large number of reactions are considered to be biologically inefficient. We thus introduce a time limitation z and pathways that can make all target metabolites producible by $t = z$. In the following, we consider the problem of finding minimal active pathways with respect to π_K and z.

V. Knockout Effects

This section provides how to predict knockout effects. In the following, we give some definitions for the prediction. Let $\pi = (N, A, M_S, M_T)$ and $\pi_K = (N, A, M_S, M_T, K)$ be an extended pathway instance and a knockout instance, respectively. In addition, we denote the number of minimal active pathways of π as $|\pi|$ and the number of minimal active pathways of π_K as $|\pi_K|$. Obviously, $|\pi_K| \leq |\pi|$ holds. Then, the gene knockout effect, i.e., the prediction by the proposed method, is given by $E_K = |\pi| - |\pi_K|$. Let K_a and K_b be sets of knockout genes. If $E_{K_a} > E_{K_b}$ holds, then we say that the gene knockout effect of K_a is stronger than that of K_b. If $|\pi_K| = 0$, i.e., $E_K = |\pi|$, then we say that the knockout effect of K is *critical* to produce target metabolites. Various metabolites are known as vital metabolites, which means organisms cannot survive without them. That is, if some gene knockouts are critical to produce such metabolites, then a given organism cannot grow any more or dies. If $|K| = 1$ and its effect is critical to produce vital metabolites, then we say that the gene $g \in K$ is *essential*.

In the following, we explain the above definition with a specific example. Suppose that we are given a pathway instance $\pi = (N, A, M_S, M_T)$, where N and A are from the extended pathway shown in Figure 2, and the source metabolite is $M_S = \{m_1\}$ and the target metabolite is $M_T = \{m_6\}$. Obviously, $|\pi| = 3$ and the minimal active pathways of π are specifically as follows: $\{r_1, r_5, r_7\}, \{r_2, r_3, r_6, r_7\}, \{r_2, r_4, r_6, r_7\}$. Then, we consider the following knockout instances π_{K_1} and π_{K_2}, where $K_1 = \{g_1\}$ and $K_2 = \{g_6\}$. For π_{K_1}, minimal active pathways including r_1 can no longer be solutions, i.e., $|\pi_{K_1}| = 2$. For π_{K_2}, minimal active pathways including r_6 can no longer be solutions either. Thus, $\{r_2, r_3, r_6, r_7\}$ and $\{r_2, r_4, r_6, r_7\}$ are deleted from the solutions of π, i.e., $|\pi_{K_2}| = 1$. Consequently, we can say that the knockout effect of K_2 is stronger than that of K_1. Moreover, suppose that $K = \{g_7\}$. Then, there is no minimal active pathway of π_K and we say that the knockout effect of K is critical to produce m_6. If m_6 is a vital metabolite, we can simultaneously say that g_7 is an essential gene.

In addition to the number of remaining minimal active pathways after knockouts, an important factor in the prediction is the gain of ATPs. This is because pathways that are inefficient with respect to energy consumption will not be used in organisms. Let $|\pi^{a+}|$, $|\pi_K^{a+}|$ be the number of minimal active pathways of π and π_K, which gain ATPs, respectively. Then, the gene knockout effect with respect to ATP production is given by $E_K^{a+} = |\pi^{a+}| - |\pi_K^{a+}|$. In particular, it is important when we consider the glycolysis pathway since one of its main functions is to gain ATPs. However, we cannot find any pathways producing ATPs on some other pathways, i.e., minimal active pathways on them must consume ATPs. In this case, the number of minimal active pathways, which consume fewer ATPs, should be considered instead of $|\pi^{a+}|$ and $|\pi_K^{a+}|$.

VI. Computational Method

This section explains how to compute $|\pi_K|$. In this paper, we use the method of computing all minimal active pathways of π proposed by Soh and Inoue [7]. This method computes pathways through propositional encoding and minimal model generation. An advantage is that this method is flexible for adding biological constraints. Moreover, we can utilize SAT technologies, which have been developed actively in recent years.

In the following, we briefly explain the propositional encoding to compute minimal active pathways of π. Let i, j be integers denoting indices for metabolites and reactions. Let t be an integer variable representing time. Let $\pi = (N, A, M_S, M_T)$ be an extended pathway instance, where $N = M \cup R \cup G$, $A = A_M \cup A_G$. We introduce two kinds of propositional variables. Let $m_{i,t}^*$ be a propositional variable, which is $true$ if a metabolite $m_i \in M$ is producible at time t. Let $r_{j,t}^*$ be a propositional variable, which is $true$ if a reaction $r_j \in R$ is activatable at time t.

The encoding of the problem of finding minimal active pathways with respect to π_K and z is as follows.

$$\psi_1 = \bigwedge_{0 \leq t < z} \bigwedge_{m_i \in M} \left(m_{i,t}^* \rightarrow m_{i,t+1}^* \right)$$

$$\psi_2 = \bigwedge_{0 \leq t < z} \bigwedge_{r_j \in R} \left(r_{j,t}^* \rightarrow r_{j,t+1}^* \right)$$

$$\psi_3 = \bigwedge_{1 \leq t \leq z} \bigwedge_{r_j \in R} \left(r_{j,t}^* \rightarrow \bigwedge_{m_i \in s(r_j)} m_{i,t-1}^* \right)$$

$$\psi_4 = \bigwedge_{1 \leq t \leq z} \bigwedge_{r_j \in R} \left(r^*_{j,t} \rightarrow \bigwedge_{m_i \in p(r_j)} m^*_{i,t} \right)$$

$$\psi_5 = \bigwedge_{m_i \in (M \setminus M_S)} \bigwedge_{1 \leq t \leq z} \left(m^*_{i,t} \rightarrow m^*_{i,t-1} \vee \bigvee_{r_j \in p'(m_i)} r^*_{j,t} \right)$$

$$\psi_6 = \bigwedge_{m_i \in M_S} m^*_{i,0} \wedge \bigwedge_{m_{i'} \in M \setminus M_S} \neg m^*_{i',0}$$

$$\psi_7 = \bigwedge_{m_i \in M_T} m^*_{i,z}$$

The formulas ψ_1 and ψ_2 represent that once a metabolite (or a reaction) is made to producible (or activatable), then it remains in the producible (or activatable) state. The formula ψ_3 represents that if a reaction r_j is activatable at time t then its reactants must be producible at time $t-1$. The formula ψ_4 represents that if a reaction r_j is activatable at time t then its products must be producible at time t. The formula ψ_5 represents that if a reaction m_i is producible then either two states hold: the metabolite m_i is producible at $t-1$ or at least one reaction r_j is activatable. The formulas ψ_6 and ψ_7 represent source metabolites and target metabolites. We denote the conjunction of $\psi_1, \ldots, \psi_7$ as Ψ_z. Then, we can enumerate minimal active pathways with respect to π_K and z by computing minimal models of Ψ_z with respect to $V^z = \{r^*_{i,z} | r_i \in R\}$.

The computation for π is always needed to compare a wild cell and its mutant. We thus explain a method to compute all minimal active pathways of π_K for a set of knockout genes K. Actually, when the minimal active pathways of π are obtained, we do not need much additional computation. All minimal active pathways of π_K are obtained by selecting pathways that do not contain some $r \in R_K$, where $R_K = \{r \in c'(g) \mid g \in K\}$. The procedure is given as follows: (i) enumerate all minimal active pathways with respect to π and z, (ii) delete minimal active pathways including some $r \in R_K$, where $R_K = \{r \in c'(g) \mid g \in K\}$. As well as the above procedure, there is another way to compute all minimal active pathways with respect to π_K and z. The same is achieved by adding constraints, which inhibit the activation of each reaction in R_K, to the formula Ψ_z.

VII. Experimental Results

This section provides experimental results. At first, we describe experimental conditions. Then, we show the results of our prediction of knockout effects for glycolysis and amino acids biosynthesis.

A. Experimental conditions

We constructed extended pathways from EcoCyc [8] and KEGG [9]. Specifically, we use EcoCyc to construct metabolic pathways, which consists of 1222 metabolites and 1920 reactions. Moreover, we use KEGG to construct relations between enzymatic reactions and genes. In the following experiments, the entire extended pathway is constructed from these two databases. Each experiment has been done using a PC (3.2GHz CPU) running on OS X 10.6. For computation, we use a SAT solver Minisat2 [10]. Koshimura *et al.* proposed a procedure computing minimal models with SAT solvers [11]. We follow their procedure to generate minimal models by using a SAT solver.

To evaluate our method, we use the Keio collection as is described in Section II. In particular, we use their results on the MOPS medium whose main nutrient is glucose. Since these comparative data are obtained from every single gene knockout, in the following, we basically consider that the set of knockout genes K consists of one gene. Moreover, in the Keio collection, if a cell growth is less than 0.1 or not applicable (N.A.) then we say that the cell is strongly affected by a gene knockout.

B. Results for Glycolysis Analysis

First, we analyze the glycolysis pathway of *E. coli*. In accordance with the MOPS medium of the Keio collection [4], we choose source metabolites as follows: β-D-glucose-6-phosphate, H^+, H_2O, ATP, ADP, phosphate, and NAD^+. In addition, pyruvate is given as the target metabolite to analyze glycolysis.

We then compute all minimal active pathways from the entire metabolic pathway of *E. coli*. As we can see in biological literature such as the work of Ferguson *et al.* [12], glycolysis is known to a pathway constructed by eight steps. However, if some reactions are disabled, then *E. coli* is expected to use other bypass pathways by additional reactions. In this experiment, we thus give $z = 12$. Moreover, the number of reactions included in each pathway is limited to less than or equal to 12.

At first, we computed all minimal active pathways using the above conditions and obtained 75 minimal active pathways. We then connected 61 genes to reactions included in them by API on KEGG. Next, we computed minimal active pathways of for each gene knockout. This experiment was done within four seconds. Figure 3 shows the results of 61 gene knockouts. The x-axis denotes each gene knockout and the y-axis denotes the number of minimal active pathways. As is shown in the figure, we compute minimal active pathways of $\pi_{K_1}, \ldots, \pi_{K_{61}}$ such that $K_1 = \{\text{b4025}\}, K_2 = \{\text{b0963}\}, \ldots, K_{61} = \{\text{b2464}\}$. However, since some of the 61 genes construct isozymes, such single gene knockout K_i does not affect the number of minimal active pathways $|\pi_{K_i}|$. However, for reference, we compute the effect of the gene knockouts that disables all of them. For instance, b2133 and b1380 construct isozymes. In this case, the number of minimal active pathways in the figure shows the case of the gene knockout of both b2133 and b1380. For each gene knockout, we computed the gain of ATP in each minimal active pathway, which is calculated by counting the number of both reactions with the coefficient of ATP: ones

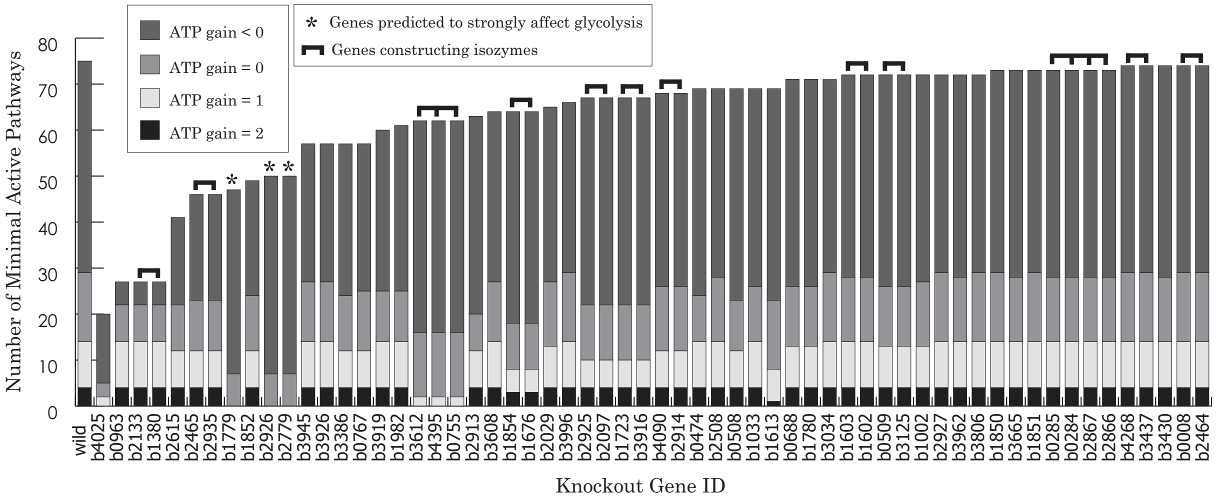

Figure 3. The Number of Minimal Active Pathways for each Gene Knockout on Glycolysis

consuming ATP and the other ones producing ATP. Minimal active pathways that produce the positive number of ATPs are more important than the others because producing ATP is a main function of glycolysis.

From the figure, we can see that *E. coli* keeps almost all minimal active pathways even by more than half of single gene knockouts. This is considered to indicate the robustness of *E. coli*. However, some gene knockouts dramatically reduce the number of minimal active pathways. In particular, the single gene knockouts of b1779, b2926, and b2779 destroy all minimal active pathways producing ATPs. Thus, they are predicted to strongly affect the glycolysis of *E. coli*.

To evaluate the above predictions, we compare them with the Keio collection. Table I compares the first 11 gene knockouts regarding the number of lost minimal active pathways. Column 1, Gene ID, shows identifiers of genes other than *wild*, which denotes an empty set of knockout genes, e.g., $K = \emptyset$. Other rows denote the result of single gene knockout. Column 2, Total, shows the total number of minimal active pathways, i.e., $|\pi_{K_i}|$. Columns 3 to 6 show the number of minimal active pathways, each of which denotes the gain of ATPs. Column 7, MOPS24hr, and Column 8, MOPS48hr, show the cell growth of *E. coli* after 24 hours and 48 hours, respectively. Note that N.A. (not applicable) refers to essential genes [4].

As the first row of Table I shows, we found 14 minimal active pathways that produce the positive number of ATPs on the wild cell of *E. coli* while there are 75 in total.

Distinguished single gene knockouts are K_8 = b1779, K_{10} = b2926, and K_{11} = b2779. Each gene knockout effect with respect to ATP production is $E^{a+}_{K_8} = E^{a+}_{K_{10}} = E^{a+}_{K_{11}} = 14$ and it is the strongest gene knockout effect with respect to ATP production, which is the important function of glycolysis. For this prediction, the Keio collection shows “N.A” for each gene knockout. Thus, in glycolysis, our predictions successfully agree with the results of the Keio collection.

Table I
11 Single Gene Knockouts for Glycolysis

Gene ID	# of Minimal Active Pathways					Keio Collection [4]	
	Total	ATP Gain				MOPS24hr	MOPS48hr
		2	1	0	<0		
wild	75	4	10	15	46	0.219-0.392	0.216-0.480
b4025	20	0	2	3	15	0.137	0.542
b0963	27	4	10	8	5	0.293	0.371
b2133[a]	27	4	10	8	5	0.303	0.366
b1380[a]	27	4	10	8	5	0.357	0.393
b2615	41	4	8	10	19	N.A.	N.A.
b2465[b]	46	4	8	11	23	0.311	0.315
b2935[b]	46	4	8	11	23	0.317	0.327
b1779*	47	0	0	7	40	N.A.	N.A.
b1852	49	4	8	12	25	0.231	0.223
b2926*	50	0	0	7	43	N.A.	N.A.
b2779*	50	0	0	7	43	N.A.	N.A.

Table II
Critical Gene Knockouts for Amino Acids Biosynthesis

Gene ID	Unsynthesized Target	Keio Collection [4]	
		MOPS24hr	MOPS48hr
wild	-	0.219-0.392	0.216-0.480
b2153	MET	N.A.	N.A.
b2615	VAL, LEU, THR, ILE, LYS, MET	N.A.	N.A.
b0004	THR	0.000	0.000
b0003	THR	0.004	0.010
b3870	TRP, MET	0.005	0.015
b2329	TRP, PHE, TRP	0.009	0.020
b2838	LYS	0.012	0.021
b3389	PHE, TRP, MET	0.010	0.032
b0074	LEU	0.026	0.034
b3177	MET	0.283	0.293
b4019	MET	0.357	0.509

On the other hand, our method predicted that there are still minimal active pathways that produce the positive number of ATPs after the single gene knockouts of b4025, b0963, and b1852. Those remaining pathways are supposed to be used as bypass pathways. For instance, b4025 encoding glucosephosphate isomerase gene of glycolysis pathway that transfer D-glucose 6-phosphate to D-fluctose 6-phosphate.

However, pentose phosphate pathway is available as a bypass pathway from D-glucose 6-phosphate, resulting in the gene knockout slow-growth at starting MOPS24hr and same level of wild cell final growth at MOPS48hr. Moreover, the knockouts of b2133, b1380, b2465, and b2935 do not affect to the cell growth since they construct isozymes.

The single gene knockout of b2615 is different to the above gene knockouts. Our method predicts that this knockout does not affect the cell growth in terms of glycolysis. However, the Keio collection shows that this is an essential gene for *E. coli*. One assumption is that it affects other functions in the cell. In relation to this, we have additional experiments for amino acid generation in Section VII-C.

C. Results for Amino Acids Generation

We also applied our prediction method to predict gene knockout effects of the cell growth in terms of amino acid biosynthesis. Since we want to involve more genes for our prediction, we particularly focus on essential amino acids for humans, whose synthesis needs more reactions than others. In the experiments, we separately constructed pathway instances, each of which consists of the following eight amino acids as a target metabolites: L-valine (VAL), L-leucine (LEU), L-phenylalanine (PHE), L-isoleucine (ILE), L-threonine (THR), L-lysine (LYS), L-tryptophan (TRP) and L-methionine (MET). In addition, to produce the above amino acids, we added the following metabolites to the source metabolites used in the glycolysis analysis: coenzyme-A and sulfite. For each of the eight amino acids, the computation time is on average 255 seconds and the longest computation time is 877 seconds.

In contrast to the result of glycolysis, we found there are 11 single gene knockouts that destroy all minimal active pathways without the limitation of z. That is, no pathway can synthesize each target on the entire metabolic pathway of *E. coli* with those single gene knockouts. Obviously, they are predicted to be critical to produce each amino acid. Table II shows the cell growth of Keio collection. Column 1, gene ID, shows knockout genes predicted as critical by our prediction. Column 2, unsynthesized target, shows target amino acids, which cannot be synthesized with the knockout of the gene in Column 1. Columns 3 and 4 show the cell growth of *E. coli* after 24 hours and 48 hours, respectively. At first, the gene knockout of b2615 is predicted as critical for the cell growth in terms of six amino acids biosynthesis. This result is also supported by the Keio collection. We thus consider the essentiality of b2615 to be caused by its knockout effect in amino acids biosynthesis rather than glycolysis. Table II also shows that our method predicts that no way to produce target metabolites with each single gene knockout: b0004, b0003, b3870, b2329, b2838, b3389, and b0074. However, the Keio collection shows that *E. coli* survives with very low cell growth. One explanation for the results is that they are suspected to keep living by consuming unsynthesized amino acids from other individual cells. In this case, since the amino acids cannot be sustainably produced, those genes are recognized to be approximately essential for *E. coli*.

Furthermore, the result of the Keio collection shows that the knockouts of b3177 and b4019 are not critical, although our method predicts them to be critical. We have detailed discussions on these gene knockouts in the following section.

VIII. Discussion and Related Work

This section provides detailed discussion about the difference of our prediction and the cell growth of the Keio collection. Figure 4 shows the glycolysis pathway obtained from KEGG [9]. Each enzyme label is replaced to its corresponding gene identifier. The figure also shows four essential genes in terms of the glycolysis pathway confirmed by the Keio collection. Our method predicted three out of four essential genes. However, b2925 is not expected to be critical for the cell growth since the gene knockout cell keeps almost all minimal active pathways that gain ATPs even if we delete both b2097 and b2925. Specifically, the knockout lost only four minimal active pathways that gain one ATP (see Figure 3). Thus, two hypotheses come up. One is that the four lost minimal active pathways are the most important pathways in glycolysis. The other is that the essentiality is caused by the breakdown of other cell functions, similar to the case of b2615. Exploring this issue is a future topic.

The difference between b3177 and b4019 in terms of amino acid biosynthesis also introduces interesting issues. At first, we consider b4019, which constructs an enzymatic reaction methionine synthase. Its conversion is as follows: 5-methyltetrahydrofolate $+$ L-homocysteine $=$ tetrahydrofolate $+$ L-methionine. In both KEGG and EcoCyc databases, there are two alternative reactions and their corresponding genes to the above reaction and b4019. A reaction S-adenosyl-L-methionine uses S-methyl-L-methionine instead of 5-methyltetrahydrofolate. On the other hand, a reaction 5-methyltetrahydropteroyltriglutamate uses 5-methyltetrahydropteroyltri-L-glutamate. However, both metabolites cannot be synthesized from the source metabolites. Specifically, S-methyl-L-methionine can be synthesized only from methionine, which is the target amino acid, and there is no reaction in the metabolic pathway of EcoCyc that can synthesize 5-methyltetrahydropteroyltri-L-glutamate. The gene b3177 is on folate biosynthesis and there is no alternative in the databases. Two hypotheses are as follows: there are unknown complementary genes, or there are unknown bypasses. For the above issues, we need to do more research on more databases and literature.

There are several researches on metabolic pathway analyses. Schuster *et al.* proposed a method called elementary mode analyses [13]. They focused on metabolic flux distributions corresponding to sets of reactions in metabolic pathways. A different point from our method is that their approach needs to define source metabolites strictly with a

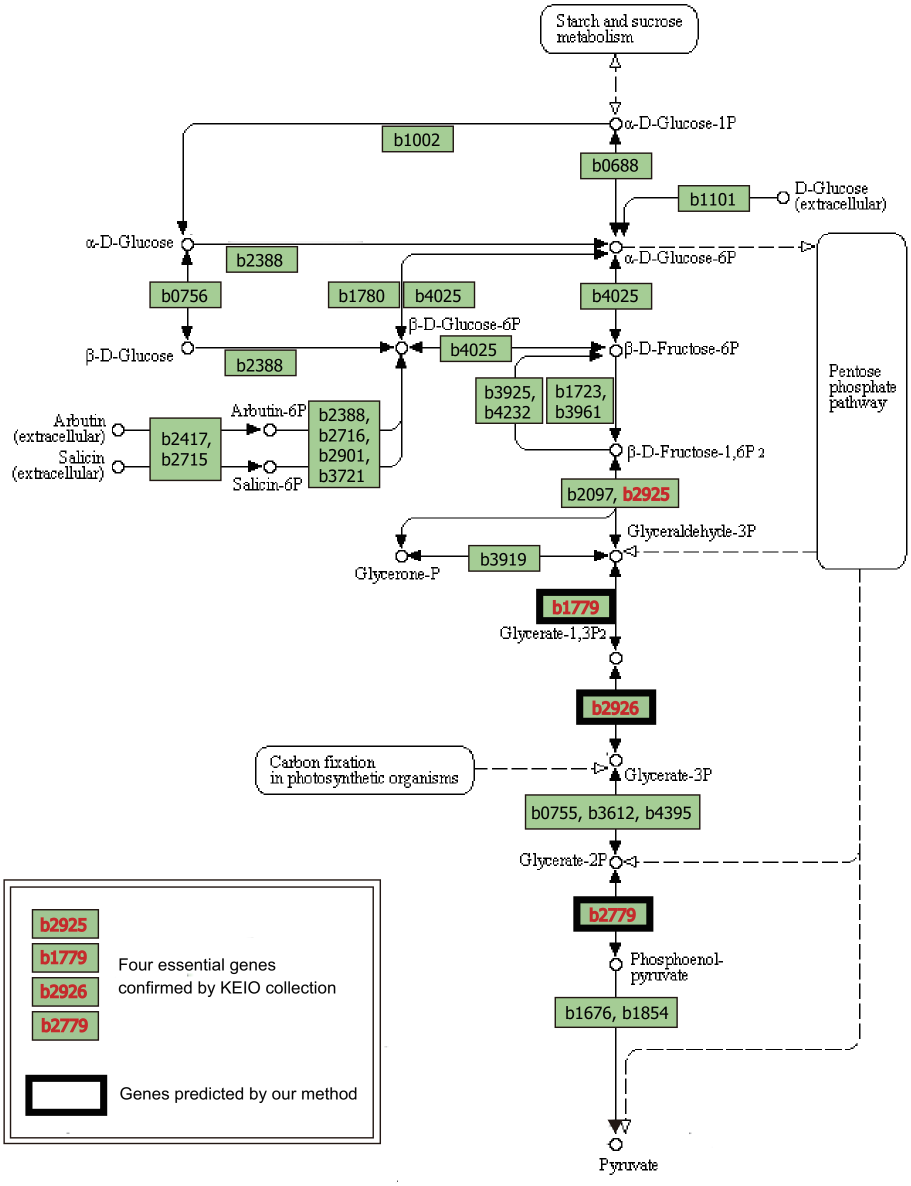

Figure 4. Glycolysis Pathway of *E. coli* K-12 MG1655 from KEGG [9]

fixed amount that must be consumed in flux. In contrast, our method treats them as candidates that will be utilized; thus, we can flexibly give source metabolites. Handorf *et al.* proposed the *inverse scope problem* [14]. This is the problem of finding necessary source metabolites from target metabolites. The two differences between their problem and our proposal are as follows. One is that they only computed the cardinality minimal solution. Unlike their approach, we can generate subset minimal solution by minimal model generation. Another one is that each of their solutions includes all reactions, which are activatable from source metabolites needed to generate target metabolites. For instance, if there are two ways to produce a metabolite from source metabolites then both are mixed in one solution, that is, we cannot distinguish between them. On the other hand, our method can distinguish between the two ways, and we think that it is important to identify functionally minimal active pathways. Schaub and Thiele applied answer set programming (ASP) to solve the inverse scope problem [15], while we use propositional encoding and minimal model generation to compute minimal active pathways.

IX. Conclusion and Future Work

In this paper, we propose a method to predict the knockout effect by enumerating minimal active pathways. We formalize the extended pathway and show the definition of minimal active pathways on it. In addition, we present a computation method for the prediction. An advantage of our method is that it allows us to trace the reason for the prediction results, e.g., we can suggest the reason for the essentiality of three genes in the glycolysis pathway. This is an important feature that other methods do not have.

In the experiments, we applied our method to extended pathways of *E. coli* and made comparisons using the Keio collection. For the prediction of the knockout of 61 genes in the glycolysis pathway, our method predicted three essential genes, which correspond to the results of the Keio collection. Moreover, we found two essential genes and nine approximately essential genes in amino acids biosynthesis. However, for the knockout of b2925, b3177, and b4019, our prediction indicated different results from the Keio collection. Revealing the reason for this difference is a future work. Moreover, we plan to evaluate the efficiency of the computation method and compare it with other methods. Although we treat relations between genes and enzymatic reactions that have one-to-one relations, we intend to extend them to relations that are more complex such as multiple relations and consider interactions among genes. Following that, we plan to apply our method to other organisms such as mice. In addition to *E. coli*, mice are well known model organisms for human study, and information available on them has been accumulated in the last decade. In particular, chromosome substitution strains are used to reveal the function of genes [16]. In addition to gene knockouts, we could adapt our method to such strains. Although there is a large difference between *E. coli* and mice, the basic metabolism is same. This fact tells us that our method can also be a potential prediction method for mice.

Acknowledgments

This research is supported in part by the 2008-2011 JSPS Grant-in-Aid for Scientific Research (A) No. 20240016, the research project of Systems Biology of Genetic Function, and the 2nd Transdisciplinary Seeds Research in the Transdisciplinary Research Integration Center. The authors would like to thank anonymous reviewers for their helpful comments.

References

[1] H. D. Jong, “Modeling and simulation of genetic regulatory systems: A literature review,” *Journal of Computational Biology*, vol. 9, pp. 67–103, 2002.

[2] M. Terzer, N. D. Maynard, M. W. Covert, and J. Stelling, “Genome-scale metabolit networks,” *Systems Biology and Medicine*, vol. 1, no. 3, pp. 285 – 297, 2009.

[3] C. J. Tomlin and J. D. Axelrod, “Biology by numbers: mathematical modelling in developmental biology,” *Nature Reviews Genetics*, vol. 8, no. 5, pp. 331 – 340, 2007.

[4] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori, “Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection,” *Molecular Systems Biology*, vol. 2, no. 2006.0008, 2006.

[5] H. Mizoguchi, H. Mori, and T. Fujio, “*Escherichia coli* minimum genome factory,” *Biotechnology and Applied Biochemistry*, vol. 46, no. 3, pp. 157–167, 2007.

[6] N. Ishii, K. Nakahigashi, T. Baba, M. Robert, T. Soga, A. Kanai, T. Hirasawa, M. Naba, K. Hirai, A. Hoque, P. Y. Ho, Y. Kakazu, K. Sugawara, S. Igarashi, S. Harada, T. Masuda, N. Sugiyama, T. Togashi, M. Hasegawa, Y. Takai, K. Yugi, K. Arakawa, N. Iwata, Y. Toya, Y. Nakayama, T. Nishioka, K. Shimizu, H. Mori, and M. Tomita, “Multiple high-throughput analyses monitor the response of *E. coli* to perturbations,” *Science*, vol. 316, no. 5824, pp. 593–597, 2007.

[7] T. Soh and K. Inoue, “Identifying necessary reactions in metabolic pathways by minimal model generation,” in *PAIS 2010, Proc. ECAI 2010*, 2010, pp. 277–282.

[8] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muiz-Rascado, C. Bonavides-Martinez, S. Paley, M. Krummenacker, T. Altman, P. Kaipa, A. Spaulding, J. Pacheco, M. Latendresse, C. Fulcher, M. Sarker, A. G. Shearer, A. Mackie, I. Paulsen, R. P. Gunsalus, and P. D. Karp, “EcoCyc: a comprehensive database of *Escherichia coli* biology,” *Nucleic Acids Research*, vol. 39, no. suppl 1, pp. D583–D590, 2011.

[9] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, “KEGG for integration and interpretation of large-scale molecular data sets,” *Nucleic Acids Research*, 2011.

[10] N. Eén and N. Sörensson, “An extensible SAT-solver,” in *Proc. the 6th International Conference on Theory and Applications of Satisfiability Testing*, 2003, pp. 502–518.

[11] M. Koshimura, H. Nabeshima, H. Fujita, and R. Hasegawa, “Minimal model generation with respect to an atom set,” in *Proc. the the 7th International Workshop on First-Order Theorem Proving*, 2009, pp. 49–59.

[12] G. P. Ferguson, S. Totemeyer, M. J. MacLean, and I. R. Booth, “Methylglyoxal production in bacteria: suicide or survival?” *Archives of Microbiology*, vol. 170, no. 4, pp. 209–218, 1998.

[13] S. Schuster, D. A. Fell, and T. Dandekar, “A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks,” *Nature Biotechnology*, vol. 18, pp. 326–332, 2000.

[14] T. Handorf, N. Christian, O. Ebenhöh, and D. Kahn, “An environmental perspective on metabolism,” *Journal of Theoretical Biology*, vol. 252, no. 3, pp. 530 – 537, 2008.

[15] T. Schaub and S. Thiele, “Metabolic network expansion with answer set programming,” in *Proc. the 25th International Conference on Logic Programming*, 2009, pp. 312–326.

[16] T. Takada, A. Mita, A. Maeno, T. Sakai, H. Shitara, Y. Kikkawa, K. Moriwaki, H. Yonekawa, and T. Shiroishi, “Mouse inter-subspecific consomic strains for genetic dissection of quantitative complex traits,” *Genome Research*, vol. 18, no. 3, pp. 500–508, 2008.